考试学

徐玖平著



成都科技大学出版社

考　试　学

徐玖平　著

成都科技大学出版社

内 容 提 要

本书是关于考试问题的一本专著。全书共十章，分别对考试的概念、作用、功能、考试设计、命题方法、考试科学性的评价、试卷分析、合格分数拟定、应考心理训练与卫生、考试与智力测验等问题进行了深入浅出的论述。全书材料丰富，内容新颖，具有较大的实用价值和一定的理论价值。适合于大、中、小学教师，教育部门管理工作者；初、高中学生、大学生、成人类高校学生阅读，特别对参加高考的学生有一定帮助，也可作为师范院校的教材，并可供教育科学工作者参考。

考 试 学

徐玖平 著

成都科技大学出版社出版
四川省新华书店经销
成都科技大学印刷厂印刷
开本:787×1092毫米 1/32 印张:10.3125
1989年6月第1版 1989年6月第1次印刷
印数：1—10000册 字数：223千字

书号：ISBN7-5616-0338-X/G·74

定价：3.60元

前　言

考试，是涉及千家万户的大事，它使多少人烦恼、焦心，又使多少人高兴、振奋！据史书记载，远在公元前二千一百多年的夏朝，我国就有了学校，考试就随学校产生而产生。可以说考试的历史已有几千年了。但是，考试有规律吗？古今中外的考试都科学吗？考试自身能成为一门独立的学科吗！回答是肯定的。正是由于我坚信这个结论，才使我撰写这本研究考试问题的专著——《考试学》。我将以此书献给：

辛勤耕耘的老师

苦心教育的家长

奋发进取的学生

作为一个教师来讲，他总希望出一套高水平的考题，总是想做好考试的组织设计工作，使考试的成绩能够真正反映学生的水平。本书用了六章的篇幅，对命题计划的编制，命题基本原则和方法、各种试题的选择和组合，进行了较详细的讨论。第三章和第七、八章则专门研究了试卷分析。评分和考试结果的可信度问题。讨论了命题的基本的原则和方法，得出了一组评价指标和获取指标值的数学表达式。长期以来，有不少教师在试题编制、试卷构成、试卷分析、评分等问题上多是凭经验办事。有些考题出得相当不错，但也有不少考题不够理想。给分标准更是不一致。这些章节有可能启发他们将经验上升为理论，尽快地掌握考试的基本规律，提高出题质量，使教师成为一个高水平的主考人。

对于望子成才的家长，他们常常受子女的考试成绩——分数所左右，对分数认识绝对化，唯分数论，不少青少年学生，因为分数问题而失去了许多受教育或就业等机会，造成了人力的浪费，压制了一些人才的成长。本书提出了考试是检查一个人知识、智力、能力的最终标准的中介，旨在能正确认识考试的意义，作用以及功能等，正确对待青少年成长的问题。做到有分数论，不唯分数论。

对于考生特别是高考生，由于各种原因考试成绩常不能体现自己的真实水平。本书专章讨论了考试的心理训练与卫生，对考生在准备考试和实际考试中，如何消除情绪过敏，心理疲劳、大脑反映迟钝、信心不足、精神紧张等心理障碍，保持最佳考试竞技状态的问题，特别是有意识进行考试心理训练的方法和注意事项，本书作了仔细的论述和介绍。对考试的生理卫生，备考和实考中应试者的营养、休息、生活规律性、最佳备考环境等问题也作了一些探索性的研究。

近年来，美国的标准化考试传到中国，引起了教育改革者的激动，特别是B·S布卢姆的认知领域学习目标分类学说，对我国考试制度和方法引起强烈的反响和震动，触发和推动了我国的考试改革。不少人用B·S布卢姆的认知目标分类来作为命题的“双目”表，被视为考试的革新。我认为这与早在唐朝时期科举取士中的“帖经”、“墨义”、“策问”、“诗赋”非常相似。当然这结论是否正确还可以讨论。但在大量“引进”外来文化的同时，挖掘中国古代文化教育遗产中的精华是非常必要的。对此进行深入的研究，以期光大中华传统文化是有利于振兴教育事业的。

本书的研究仍是非常肤浅的，有待于同仁们的进一步充

实与探讨，仅供广大读者选择、思考、批评、研究。如果本书能给予读者点滴的启发，能有助于教育质量的提高、人才的选拔，那正是我所期待的。由于我受所掌握的资料的限制，本书所涉及的调查资料的地理分布也不够均衡。如有机会修订这本书，我将尽可能在《考试学》中涉及各国、各地、各民族有代表性的考试方法。

本书是从1985年秋开始写作的，1987年夏于重庆西南师范大学梅园完稿。在艰苦的撰写过程中曾得到不少朋友和同仁们的热情支持与帮助。特别是海外朋友们的鼓励，在出版工作中，成都科技大学陈正权提出了不少建设性的意见，并在百忙中抽出时间来做该书的责任编辑，在此一并致以谢忱。我还要说明的是：由于我水平有限，此书无论在体系上、材料引用上以及具体内容的论述上都可能有疏漏和错误，尚祈专家学者们不吝赐教！

作者于成都

目　录

第一章　总论……………………………………（1）
第一节　考试的概念……………………………（1）
一、考试的定义……………………………（1）
二、考试与教育评价………………………（3）
三、考试与教育评估………………………（4）
四、考试与教育测量………………………（6）
第二节　考试的历史发展………………………（8）
一、古代学校考试…………………………（8）
二、古代科举考试…………………………（11）
三、唐代考试方法与B·S布卢姆的认知目标分类的思考…………………………（16）
四、现代考试的特点………………………（22）
第三节　考试学的任务和内容体系……………（24）
一、考试学的任务…………………………（24）
二、研究考试学的意义……………………（25）
三、研究考试学的根据……………………（27）
四、考试学的基本体系和主要内容………（29）
第四节　考试学的基本特征……………………（29）
一、实践性…………………………………（31）
二、科学性…………………………………（31）
三、整体性…………………………………（32）

四、动态性……………………………………（32）
五、超前性……………………………………（33）
第五节　考试学的研究方法……………………（33）
一、经验方法…………………………………（35）
二、理论方法…………………………………（35）
三、思维方法…………………………………（36）
四、系统方法…………………………………（37）
五、数学方法…………………………………（38）

第二章　考试的作用、测量特点、功能与设计……………………………（40）

第一节　考试的作用…………………………（40）
一、考试的目的………………………………（40）
二、考试的作用………………………………（42）
三、考试的局限………………………………（44）
第二节　考试与测量…………………………（46）
一、测量的概念………………………………（46）
二、人的知识、智力、技能的可测性………（48）
三、考试是一种特别的测量…………………（51）
第三节　考试的功能…………………………（52）
一、系统的观点………………………………（52）
二、信息的观点………………………………（53）
三、控制的观点………………………………（54）
第四节　考试目标、内容和考试标准的制定……（58）
一、考试目标的规定…………………………（58）

二、考试内容的确定…………………………（59）
三、考试标准的制定…………………………（60）
第五节　考试方法和类型的选择………………（61）
一、常用的考试方法…………………………（61）
二、常用的考试类型…………………………（63）
三、考试方法和类型的选择…………………（64）

第三章　考试科学性的评价指标………（66）

第一节　考试的信度……………………………（66）
一、信度的概念………………………………（66）
二、信度的种类与计算方法…………………（69）
三、提高信度的方法…………………………（78）
第二节　考试的效度……………………………（81）
一、效度的概念………………………………（81）
二、效度的种类与计算方法…………………（83）
三、提高效度的途径…………………………（87）
第三节　考试的难度与区分度…………………（89）
一、难度的概念与计算方法…………………（89）
二、区分度的概念与计算方法………………（91）
三、难度与区分度的关系……………………（94）
四、提高考试难度与区分度的途径…………（97）

第四章　命题总述……………………………（99）

第一节　命题计划的编制………………………（99）
一、编制命题计划的目的……………………（99）

二、命题计划的编制……………………………(100)
三、命题大纲举例……………………………(101)
四、认知目标分类命题…………………………(103)
第二节　命题的基本原则……………………………(106)
一、命题工作的重要性…………………………(106)
二、命题的基本原则……………………………(107)
三、命题的任务………………………………(109)
第三节　预试概述………………………………(110)
一、预试的意义………………………………(110)
二、预试的组织要求……………………………(110)
三、预试结果的分析……………………………(111)
第四节　题库的建立………………………………(112)
一、题库概述…………………………………(112)
二、建立题库的意义……………………………(113)
三、如何建立题库……………………………(114)
四、题库充实与更新……………………………(116)
第五章　主观性试题的命题……………… (117)
第一节　主观性试题概述……………………………(117)
一、主观性试题的概念…………………………(117)
二、主观性试题的优缺点………………………(119)
三、主观性试题的选用…………………………(122)
第二节　简答题的编制……………………………(123)
一、简答题的特点……………………………(123)
二、简答题的编制……………………………(124)
三、简答题的改善……………………………(125)

第三节　论述题的编制……………………………(126)
一、论述题的特点……………………………(126)
二、论述题的编制……………………………(127)
三、论述题的改善……………………………(129)
第四节　作文题的编制……………………………(130)
一、作文题的特点……………………………(130)
二、作文题的编制……………………………(132)
三、作文题的改善……………………………(134)
第六章　客观性试题的命题……………(136)
第一节　客观性试题概述……………………………(136)
一、客观性试题的概念……………………………(136)
二、客观性试题的优缺点……………………………(140)
三、主观性试题与客观性试题的比较………(141)
第二节　是非判断题的编制……………………………(143)
一、是非判断题的特点……………………………(143)
二、是非判断题的编制……………………………(144)
第三节　选择题的编制……………………………(145)
一、选择题的特点……………………………(145)
二、选择题的编制……………………………(147)
三、多项选择题的编制……………………………(149)
四、选择题的修改……………………………(151)
第四节　其它客观性试题的编制………………(153)
一、填充题……………………………(153)
二、分析判断题……………………………(154)
三、改错题，分类题、配对题………………(155)

第五节　客观性试题试卷的编制……………………(158)
一、客观性考试的种类……………………………(158)
二、客观性试题试卷的编制………………………(160)
三、客观性考试的改善……………………………(165)
第七章　分数的衍化与合格分数的拟定
……………………………………(168)
第一节　分数的衍化……………………………………(168)
一、分数衍化的意义………………………………(168)
二、考试分数的分布………………………………(170)
三、百分比值分……………………………………(172)
四、位置百分………………………………………(173)
五、标准分…………………………………………(175)
第二节　选择题评分问题的研究……………………(179)
一、猜测分数的矫正………………………………(179)
二、"轮盘赌"的思考……………………………(183)
三、时间参数………………………………………(184)
第三节　合格分数的拟定……………………………(186)
一、聂刁思基（Nedelsbi）评核法…………(187)
二、"边界组"评核法……………………………(191)
第八章　试卷分析与研究………………………(200)
第一节　试卷分析……………………………………(200)
一、试卷分析的意义………………………………(200)
二、试卷的定性分析………………………………(201)
三、统计分析………………………………………(202)

四、实例及其简单的讨论……………………(203)
第二节 "SPEI"图表分析法……………………(215)
一、"SPEI"图表的制作方法……………(215)
二、"SPEI"图表的基本性质……………(218)
三、"SPEI"图表分析的作用……………(222)
第九章 考试与智力测验………………(224)
第一节 智力的概述………………………(224)
一、智力的定义…………………………(224)
二、智力的结构…………………………(226)
第二节 考试与智力发展的辩证关系……………(231)
一、考试与智力发展的内在联系……………(232)
二、考试与智力发展的关系………………(233)
第三节 智力测验…………………………(236)
一、考试与测验…………………………(236)
二、智力测验的兴趣……………………(237)
三、智力的分配…………………………(240)
四、创造能力的测验……………………(241)
第四节 智力测验理论的主要流派……………(244)
一、心理测量学理论……………………(244)
二、皮亚杰理论…………………………(246)
三、资讯传处理论………………………(248)
第五节 智力测验结果的应用………………(252)
一、学习标准的确立……………………(252)
二、教育效果的诊断……………………(252)
三、学习评价的标准……………………(252)

四、应用于班级的编排……………………(253)
五、发现特殊儿童……………………………(253)
六、应用于升学、就职指导………………(253)
七、应用于性格观察………………………(254)
第十章　考试的心理训练与卫生…………(256)
第一节　心理训练概述……………………(256)
一、心理训练的概念………………………(256)
二、心理训练的地位与作用………………(257)
三、心理训练的分类与任务………………(260)
四、心理训练的原则………………………(261)
第二节　心理训练的基础…………………(264)
一、心理训练的气功基础…………………(264)
二、心理训练的实验心理学基础…………(266)
第三节　应试者的心理概述………………(268)
一、应试者的最佳考试状态………………(271)
二、应试者的最佳心理状态………………(272)
三、应试者最佳心理状态形成的基本因素…(273)
第四节　考试心理训练的基本方法………(278)
一、一般心理训练方法……………………(278)
二、准备具体考试的心理训练方法………(280)
第五节　考试卫生…………………………(282)
一、生理卫生………………………………(282)
二、心理卫生………………………………(284)
三、环境卫生………………………………(285)
附录：特殊型考试方法探讨………………(286)

第一章　总论

第一节　考试的概念

一、考试的概念

考试的形成伴随学校教育的产生而产生。从古代开始迄今已有几千年，考试仍不失其为检查教学、了解学生学习情况、区分考生的知识水平与智力差异的较好手段。尽管各个时代的教育目的、制度、内容、方法等方面不同，若干年来多经变革，教育从低级到比较高级，从原始到比较现代化，但考试仍被人们视为办学校的常规手段之一。然而，究竟什么是考试，这还是一个带有争论的问题。

人们常从不同角度对考试的概念作出解释。从行政管理角度看，考试是一种测量教学成绩的制度。从应用角度来看，考试是检验、评价学业水准和才智的一种手段或方法。如果从教学方面看问题，它则常被视为教学过程中的一个重要环节。那么，什么叫考试呢？考试是一种严格而又庄严的科学鉴别方法，是让考试对象在规定的时间内按指定的方式解答精心选定的题目，对解答的结果评等判分，为主考者提供考试对象某方面的知识和能力状况的信息。有着振奋精神，激发进取心的积极意义。构成考试有五个要素：其一，考试是在有人监督的情况下进行的。每个受试者都必须独立完成项

目，从而将个体从集体中严格分离出来，单独进行鉴别。其二，考试给予每个受试者以同等的条件——同样的试题，同样的考试时间，同样的考试环境，同样的评分“标准”，从而使个体的考试结果具有可比性。其三，在较正规的考试中，试卷上通常不写受试者的姓名，而以号码作为识别标志，从而避免了由于评分者与受试者之间存在某种特殊关系而对评分产生影响，保证了评分的客观性。其四，主试部门可以按照不同的考试目的设计考试的内容、难度和方式。因此，考试作为一种鉴别技术来说具有可控性。其五，考试中具有竞争性，许多人在 考场上能够象运动员在技场上那样发挥出最大的潜力。无论是青年学生或是成年人，在考试前大多有担心、紧张、兴奋、激昂、自信和期待的复杂心情。考试之后又都急不可待地想知道分数，同时又怕知道分数。大多数胜者意气风发，信心倍增；败者也有败而不馁、再求一搏的精神状态。换言之，凡是同时满足上述个体分离原则、可比性、客观性、可控性和竞争性的个人知识能力测试场合可称为考试。反之凡是考试必须同时具备以上五个基本要素。

考试已经成为社会的一个重要因素。其科学定义还有待进一步探索。但是，这并不妨碍我们从事考试研究工作。老子在《道德经》中，第一章就写道：“道可道，非常道，名可名，非常名。”老子认为，可以用言词表达的道，就不是常道，可以说得出的名，就不是常名[1]。老子的这一思想已被广泛接受。贝尔纳写道：“《道德经》，这部描

①陈鼓应《老子注释及评价》，中华书局，1984年，第62页。

写中国人对自然与社会运动看法的 中国古典优秀著作，一开始就明确告诫人们，过于刻板的定义有使精神实质被阉割的危险。[①]”我们在研究考试时，也应当遵照这种思想方法，不要 局限在下一个简单的定义。

二、考试与教育评价

评价（evaluate）即评判价值，就其内涵来说，它与价值这一概念有关。在英语中，评价这一词由词干“valu”加上词头“e”和动词性词尾“ate”组成，其中“valu”意为价值，词头“e”涵义等同于“out”意为引出。评价即为引出和阐发价值。从本质上说，评价是一价值判断的过程。

教育评价是对教育的社会价值作出判断的过程。教育作为一种客观的社会活动，它的结果满足一定社会的政治、经济、文化等发展的需要，其现实性和可能性构成了教育的政治价值、经济价值和文化价值。这些价值的总和构成了教育的社会价值。教育价值就是对这社会价值作出判断，从而推动教育活动的发展。从教育理论产生那天起，便有关于教育效果的评价。原始教育中的师带徒、父传子，师、父在传授经验和技能时，总要不时地让徒、子复述和操作，不对就予以纠正。在认为徒、子已掌握时，再传授新的经验和技能。这复述和操作，既是传授经验的过程，又是对传授结果的评价过程，还可以说这就是考试的启蒙。随着教育的发展，特别是学校教育的出现和发展，早期关于教学效果的粗糙、零

---

①J. D. 贝尔纳，A. L. 马凯“在通向科学的道路上”《科学的社会功能》，商务印书馆，1982年。

碎的评价，就被更准确、更客观地评价教育效果的方法所取代。但是，教育效果的评价是人的内部素质的提高差的度量，毕竟不同于物体外部特征的变化，度量起来是困难的。聪明的人们在实践中逐步摸索，创造评价方法，从要求教育对象掌握的教学内容中选取出部分内容（在统计学中叫做“抽样”），并编制成便于回答的问题，待教育对象按要求回答后，再对回答结果进行评等或判分。

事实上，考试产生于教育评价，并大大推动了教育评价的发展。它们之间有着十分密切和不可分割的关系。但是，教育评价与考试并不是一回事，考试只是教育评价的一种方法，是一种对受教育者的个体进行定量评价的方法。教育评价还有其他多种方法。它们并不能被考试所代替，考试在教育评价中产生之后，它作为一种测量人的知识水平和能力差异的手段，广泛应用于社会生活的多方面。它已经不单单是一种教育评价的方法了。

三、考试与教育评估

何谓教育评估，它的完整的科学概念怎样？一直是教育界、学术界争论不休的问题，在教育评估理论形成和发展的历史过程中，多次的争论不仅使教育评估的性质、功能等发生了变化，同时，也使各种观点大量涌现，直到今天国内外对教育评估的概念还未取得一致的意见。教育评估之父泰勒（〔美〕R·W·Tyler）认为：教育评估就是衡量实际活动达到教育目标的程度，或者讲它是一种对教育活动达到教育目标程度的判断。美国评估学者豪斯（E·R·House）认为：教育评估是既有描述、又有判断的活动，是一种对优

缺点和价值的评估。美国另一位学者斯塔弗尔比姆（D·L·Stuffbebean）则认为：教育评估是为决策提供有用信息的过程。教育评估工作在我国虽然只是刚刚起步，但不少学者对教育评估的概念也作了许多研究，有人认为：教育评估是一种新的教育成绩的考查方法。也有的人认为：教育评估是对教育目标和教育的社会价值进行判断的过程，目的在于为改进工作提供信息。有的信则认为：教育评估是一种信息反映，概括地说，就是根据教育的宗旨和科学的评价标准，运用现代化评价技术手段，评判估价全部教育效果的社会价值，在实践中，人们还常把教育评估看成是对教育活动或现象进行量上的估价，带有模糊性，综合性特点的评价。因此，一般认为教育评估就是定量的模糊的和综合的教育评价。

我们说，教育评估与教育评价在实际工作中很难加以严格区别，乃是由于评估与评价两个概念相近所致。评估（assessment）是根据一定的目标，通过系统地收集信息，对客体作出价值判断的过程。评价（evalute）指衡量和评定人物或事物的价值。虽然，两者都与衡量、判断客体的价值不楞分割。评估作为价值判断的过程，它的特点价值的特点有关，因此，这里我们有必要对价值这一概念略作探讨。

价值在哲学界至今仍是一众说纷云的概念。在西方哲学界，有截然对立的两种观点：一方为客观主义，它们认为，价值是客观对象固有的本性，是独立于人类的认识，情感和行为之外的；另一方是相对主义，它不同意客观主义的这种价值论，不同意关于价值是客体固有的本性的观点。它认为价值乃是用来表达个人对事物与客体的感情的。现代存在主义哲学则进一步主张，价值乃是个体创造的。这样，在他们

看来，价值是完全主观的，相对的东西。事实上，这两种观点都是错误的。按照马克思主义的观点来看，价值是一个用以表明客体对主体特殊效用关系的概念。‘价值’这个普遍的概念是从人们对待满足他们需要的外界物的关系中产生的。[①]”价值是主体人的需要同外部世界的一种关系。价值表示在实践的认识活动中，客体的存在，属性和合乎规律的运动变化结果向主体接近的现实性和可能性。价值是一以主体的一定需要、意图、愿望、一定的目的、目标、指向为准绳来衡量客体效应的。因此，价值是主体性与客体性的统一，价值的这一特点决定了作为价值判断过程的评估区别于其它活动的一个重要特征，是主体性与客体性的高度统一的活动，这是评估的基本特点。

诚然，评估作为一种价值判断的活动，它又不同于其它形式的价值判断的活动，1981年，代表着美国十二个与教育评估有关组织的十七名成员在对教育评估下定义时，提出评估乃是“对教育目标及它们价值的判断的系统调查，是为教育决策提供依据的过程。[②]”这就指出了评估是通过系统地收集信息，按照严格的科学程序，有计划，有组织地进行的活动。因此，它是关于对象的一种较为深刻的、对于它的发展变化具有重要影响的价值判断过程。

综上所述，考试是教育评估的一重要手段和方法，但是，二者绝不相等，它们各自有自己的独立的内涵与外廷。其研

①《马克思恩格斯全集》第19卷，第406页。

②Nevo,D“The Conceptualization of edeccational Evaluation: An Analytical Review of the Leterature,”Review of Edeccational Research (Spring 1983)

究的领域、方法、手段、原则均有差异。决不能误将考试代替教育评估。也决不能把教育评估视为考试。

四、考试与教育测量

测量是依据一定的标准对事物的某一属性作出事实判断，进行赋值的过程。测量是考试的基础，但是测量并不就是考试。

首先，它与考试相比，我们可以说测量是一纯客观的过程。客观性是测量质量的首要指标。客观的测量期望测量者能尽可能地排除各种主观因素的影响。不同的测量者，对同一事物的测量，除了允许的误差外，它的结果应该是相同的。考试是在测量的基础上进行的。但它却超出了描述，而试图去确定行为的价值。对同一事物，由于考试者的价值观不同，考试的结果有时可能大相径庭。在本质上，考试并非是纯客观的，而同时具有客观性和主体性这两个基本属性。它的客观性就是它对教育工作状况和结果的客观测量、客观的根据、客观的描述的基础上进行的。它的主体性就是它又是以主体的需要和目标为准绳来评判的。

其次，与考试相比，测量是一相对地说较为单一，即只重视量的获得的活动。而考试则是一个较为复杂、包含着多重活动的过程。考试，是对客体质的差异作出判断的过程。

考试与测量是有紧密联系的，任何考试都把测量作为基础，是测量的深化和发展。从教育的历史演变看，考试也是在教育测量的基础上发展起来的。历史的分析与逻辑的分析结论是一致的。但是，如上所述，它们之间是有着重要的区别的。

测量应是一客观的事实判断过程。但在实际工作中，测量者的价值观念，往往也会对测量产生重要的影响。哥德曾说守："我们见到的只是我们知道的。"这就是说对同一对象，不同的观察者有时也会出现不同的观察结果，除了难免的误差外，观察者的观念不同是一个重要的因素，它对观察结果的获得产生重要的影响，这一点上，它与考试是难以区分的。这是需要我们在工作中高度重视的地方。

第二节　考试的历史发展

一、古代学校考试

根据教育史材料提供的信息，考试的渊源起于中国，世界上最早的考试活动也是中国。在古代的学校中就建立了定期对学生进行考试的制度，并在实践中创造了许多今天仍在采用的考试方法。

在战国时期问世的《礼记·学记》中，就已经谈到考试的问题。所谓的"比年入学，中年考校，一年视离经辨志；三年视敬业乐群；五年视博习亲师；七年视论学取友；谓之小成。九年知类通达，强立而不反，谓之大成，夫然后是以化民易俗。近者说服而远者怀之，此大学之道也。"这就是明证。即是说学生入学，每隔一年必须考一次，第一年，考查他们能否分析经义，辨别志愿而决定学习趋向；第三年他们能否专心学习，与同学互相切磋研讨。第五、七年考查他们能否研讨学业之得失与识别他人的贤否，并选择善者而为友。如能达到标准，叫做"小成"。第九年，考查他们能否

推理论事，触类旁通。有否坚定不移的志愿而不再有违反师长教诲的地方。如能达到标准，叫做“大成”。

汉朝的太学，设有严格的考试制度。汉武帝初设太学时，就规定“一年辄课”制（即一年考一次）；到东汉桓帝后，又改为“二岁一试”制。太学生没有肆业年限的规定。只要通过考试就可以毕业。并按考试成绩的高低授予不同的官职。考试的方法，有“口试”，“策试”，“射策”三种。射策，就是事先出好多道试题并分别抄录下来，由学生随机抽取作答。

唐初的统治者重视兴办学校，使官学和私学并举以培植人才，学校成为人才的重要来源。根据政府法令的规定，学校中的考试比较频繁，要求定时项进行考试。归崇敬在《辟雍议》中提到考试办法时曾说“旬省月试，时考岁贡。以生徒及弟多少，为博士考课上下。”把学生考试成绩作为学官考勤的重要标准。学官当先重视考试。规定的按时考试有旬试，月试，季试，岁试。

旬试：每旬进行一次考试，唐代的学校与政府机关一样，每旬放休，休假一日。放假前一日，学校进行考试，由负责讲授的博士主持，考试这一旬内讲习的内容，采取方式是帖经和问义。据《新唐书·选举志》载：“前假，博士考试，读者千言赋一帖，帖之言，讲者二千言问大义一条，总三条，通二条为第。”这种经常性的考试，表明在教学上重视及时复习巩固，其要求偏重在记诵。

月试：每月终第三次旬试时，试一月内讲习的内容，这就是月试。起初，学校的旬试和月试配合进行，学生和博士的生活都围着考试转，显得十分紧张。对教学活动是

起一定的督促作用。久而久之，管理放松，博士和学生都感到频繁的旬试是一种过分的精神负担，需要精简。因而放弃旬试，保留月试。元和元年（806年）规定国学“每月一度试[1]”，有的专科学校根据教学特点，没有规定旬试，只规定月试，如医学，只规定“博士月一试[2]”。根据实行的情况来看，月试比较切实可行。所以月试成为固定的一种考试。

季试，每季将终时，总一季学业举行考试，称为季试。凡附国子监修业的士人，也参加每季一试。如会昌五年（845）规定：“士人修明经进士业，并隶名太学。每季一试，使经艺习熟。”季试比月试更重要些，所以有的学校由部门的领导人来主持考试，如医学就规定：“太医令丞季一试。[3]”

岁试，每年终，总一年的学业进行考试。岁试不仅是一年课业的总检查，而且考试成绩就作为升留的依据。因它比较重要，规定也更具体些。《新唐书·选举志》：“岁终，通一年之业，口问大义十条，通八条为上，六为中，五为下。”岁试通常由主管部门的领导人亲自出来主持，《新唐书·百官志》规定：国子监“丞一人……掌判监事，每岁，七学生业成，与司业、祭酒灌试，登第者上于礼部”。《唐六典·太医署》规定：医学由“太常丞年终总试”。领导人主持的考试，考试成绩直接得到认可，登第的人就可以上报，获得参加国家科举考试的资格。

学校还用考试的办法来加强学生的学籍管理，对考试优异的学生给予物资奖励和精神奖励。旬试、月试、季试的分

---

①《册府元龟．学校部》

②③《唐六典．太医署》

别给予奖励，而且把岁试奖励与它们联系在一起。当然岁试奖励为最高。如果岁试业成，不要求出仕，而愿意留在学校继续学习的，可以在学校得到升转，“诸学生通二经，俊士通三经已及第而愿留者，四门学生补太学生，太学生补国子学”。用提高等级地位，来奖励学生。对考试不合格的学生要给予惩罚和留级。如果岁试不合格，一年二年还允许留学，若连续三年不合格则采取较重的处分，“并三下与在学九岁，律生六岁，不堪贡者，罢归。”到了元和年代，还出现让考试不合格的学生停厨，使学生们在饿饭的压力下，不能不努力上课，以求得考试合格。

学校与科举两者相辅而行，学校培养人才供科举选拔，是科举赖以发展的 基础；而科举是学校生员必经的出路，成为支配学校的重要力量。唐宋以后，随着政治形势的发展，封建统治阶级越来越重视科举的作用。抓住科举作为关键一环，以左右学校，学校也就完全被纳入科举的轨道，成为科举的附庸，渐渐地成断为科举考试的预备场所了。学校被迫跟着徉举走，而落为科举的附庸。学校教育的内容方法，都必须适应科举的需要，在考试制度方面更是如此，学校无独立的特殊的考试方式，其考试方法，完全仿照科举考试，两者在精神实质上是统一的。当然一千多年来，学校的考试还是有些改进和发展的。这里该特别提及的是，宋朝时学校考试中就有积分法。到元、明时就形成了比较完备的积分制。如明朝的国子监规定，在高年级的孟、仲、季月各考试一次，每次试卷分三等。文理双优给一分，理优文劣给半分，文理俱劣的无分。在一年内积分达到八分的为及格，予以毕业；成绩优异者，经皇帝批准可提前毕业，破格录用。

二、古代科举考试

在历史上，考试不仅在教育中为培养人才起着重要的作用，成为教育中的独立的环节，而且在人才的选拔中也发挥着重要的作用。

科举制度，是隋以后各封建王朝设科考试选拔官吏的制度。由于分科取士而得名。隋文帝废除为世族垄断的九品中正制，于开皇七年（公元587）设志，行修谨，清平干（翰）济二科。炀帝时始置进士科。唐代于进士科外，又置秀才明法、明书，明算诸科，又有一史，三史，开元礼，童子，道举等科。武则天实行殿试，并增设武举。其由皇帝特诏举行者称制科。诸科之中，惟进士科为皇帝设，最为重要。宋以后科举均用经义。明清两朝以《四书》、《五经》文句为题，规定文章格式为八股文，解释须依朱熹《四书集注》等书。清光绪三十一年（1905年）推行学校教育，科举制度即废除。

中国封建社会的科举考试制度是很完备的。在实行科举考试制度的漫长时间里，它成为中国封建社会选士和任命官吏的一个主要途径。因此，它对于当时社会的进步与倒退，不能不发生重大的影响。隋唐以来，中国封建社会的许多重大问题，其中包括科学和生产的发展与落后，特别是教育，都与实行科举制度存在着直接或间接的联系。可以这样讲，隋代，从隋炀帝二年至隋灭亡的了几十年间，是科举制度的创建时期；唐代，是科举制度完备、兴盛时期；明代，科举制度开始走向没落阶段。科举制度在封建社会兴盛时期，曾发挥了促进社会发展的进步作用，随着封建社会的衰落，这个制度逐渐变得腐朽，阻碍社会的前进和发展。

科举制度是促进还是阻碍社会前进与发展，这主要看科举制度到底能不能选拔人才？不同的历史时期，科举制度起着不同的作用，“唐宋八大家”中，除苏询外，其余七人都中过进士，唐宋文学上有成就的人，除李白等少数人外，大多数人都是进士出身，其中在科举考试中名列前茅者亦不在少数。但是，历史上名人拒考或者累试不第的，也不乏其人，特别是明清两代。因此，对科举制度究竟能否选拔人才的问题，不能一概而论，而应该从各代科举取士的实际情况出发，作具体的分析。

首先，任何一种选拔人才的考试，都要规定选拔的对象，科举考试也有其的对象，这就有一个才路的宽和窄的问题。科举考试同“九品中正”制比较，所以能够选拔一部分人才，一个重要因素是它比“九品中正”制的才路宽。“九品中正”制是着眼于门第，以豪世为主。它的等级，即是以豪世为标准划分的。按照豪世不同，把地主阶级分为九个等级，前三品为上品，大小中正官均由世族豪门担任，四品以下为下品，从庶族地主中选出。从而形成了“上品无寒门，下品无世族[①]”的世族豪门世世代代做大官的局面，而科举考试的大门，不仅向世族豪门敞开着，也向庶族地主敞开着。布衣寒门子弟，符合条件的都可以参加科举考试。这是有利于选拔人才的。历史上，出身寒微而由科举做官并显后名的实例不胜枚举。著名政治家、文学家范仲俺就是一例。

各代科举考试所限定的应试对象并不完全一样。如唐代武官，会昌年间及明代，就限定科举应试对象必须由学校出

①《文献通考．选举》

身，非学校出身者不准应试。在取士有定额的情况下，限制一些人应试，可以减少科举考试中的矛盾。但是，从选拔人才的角度看，限制一部分人应考，会使才路变窄，不如才路宽些利于选拔人才。当然，不是每个朝代都限定学校出身者才准应试。也有只问学历、不限制是否由学校出身、凡读书人皆可应试的时候。从那时的科举考试中举的 情况来看，非学校出身而参加科举并且及第的，确有一部分人是人才。太平天国曾规定，参加科举考试，不受性别限制，妇女和男人同等待遇[①]。但是，在封建社会中，没有，也不可能彻底解决才路方面的问题。《镜花缘》中描绘的妇女参加科举考试的情景，不过是一种幻想而已。

其次，由于取士是科举考试制度的一个最关键、最重要、最复杂的环节，它最集中体现了封建统治阶级的政治需要。科举考试取士的标准主要是封建地主阶级的政治标准。各代科举取士的作法，常常是不一样的。它与当政者的用人路线是紧紧地联系在一起的，对于同一个人，常常有截然不同的评价。譬如，陆游在秦桧当权时参加科举考试，被秦桧的孙子秦埙顶替了。十年后，孝宗皇帝执政，却亲自召见了他，并赐为进士。尔后，当朝庭与金人议和时，陆游又遭冷遇。

唐代科举取士，某些方面一定程度上做到了考试与推荐相结合，考卷与行卷相结合。大诗人白居易赴长安应考时，向当时名人顾况行卷，行卷第一篇是《赋得古原草送行》；见了白居易的名字顾况有些看不起说："米价方贵，长安居大不易。"等他看了白居易的诗，读到"野火烧不尽，春风

①舒新城编的《中国近代教育史资料》（上）

吹又生"时，不禁拍案叫绝。说"能做这样的诗，居亦易矣"。从此，他到处赞扬白居易的才华。到清代科举取士时，名曰实行"大挑选"办法，实际上搞了重重清规戒律，科举的桎梏把人束缚得死死的。泛滥于清代的八股，则是一种害人的东西。习之者完全是为了科举进身，"一旦获偿以愿，即复弃去，消之者谓等于敲门砖[①]。"它诱使一代代的知识分子废经弃史，步人愚昧的深渊。在会试前后还要挑选，主要标准是形，貌和所谓应应对，形貌按同、田、贯、日、气、甲、由、申八字把人的形貌分为八类。而分体正为"同"；举止端疑为"田"；体貌欣长为"贯"；"骨胳精干"，为"日"，此四类为合格。形相不正为"气"；上寅下削为"甲"；上窄下粗为"由"；上下尖而中粗为"申"，此四类为合格[②]，这就是当时所谓"人文并选"，"身言并试"的大挑选。科举取士，到这种地步，已经走进了死胡同。

再次，考试的科目与内容有关，一种选拔人材的考试，应给英雄以用武之地。通过考试，把人才的本事考出来、而科举制度就难以做到，作为为封建地主阶级服务的科举制度，其考试科目和内容的确定，只能从封建地主阶级的政治需要出发。本来科目，有"文"、"理"、关系问题。封建统治者实行科举制度的目的是牢笼人士，授以官爵，为巩固其政权服务。而不是发展科学文化。在他们看来，对于巩固其统治，理科远非文科更直接。因而，科举考试重文废理的弊病十分突出。从形式上看，科举考试设科很多，似乎是文理兼有，实际上各类科目，尤其是数理科，不为士人所趣，形同

---

①②商衍鎏《古代科举考试述录》

虚设，历史上，实行科举制度期间，在自然科学方面有成就的，如数学家、医学家、天文学家、地质学家等，许多人都不是在科举中得到功名的。这种现象，主要与考试的性质和内容有关。科举考试并不是考其所长，而是考其所短。有些名人及第，其所长，与科举考试的科目和内容没有什么关系，他们在科举中得功名，如歪打正着，对他们来说，科举只不过是晋升之路。因此，对科举考试的设科，内容与人才问题，必须从实际出发作具体分析。

第四：科举考试，由于竞之者众取之者寡，及第者可以被任官，飞黄腾达，通过考试的形式来决定人的前途命运。各种势力的激烈角逐是不可避免的。于是，便产生了"温卷"、"求知己"、考试作弊的问题。在舞弊丛生的封建社会里，科举考试自然成为营私舞弊的一个重要市场。舞弊之风不仅是选拔人才的障碍，也是考试制度本身的一个主要威胁。

三、唐代考试方法与B·S·布卢姆的认识目标分类的思考

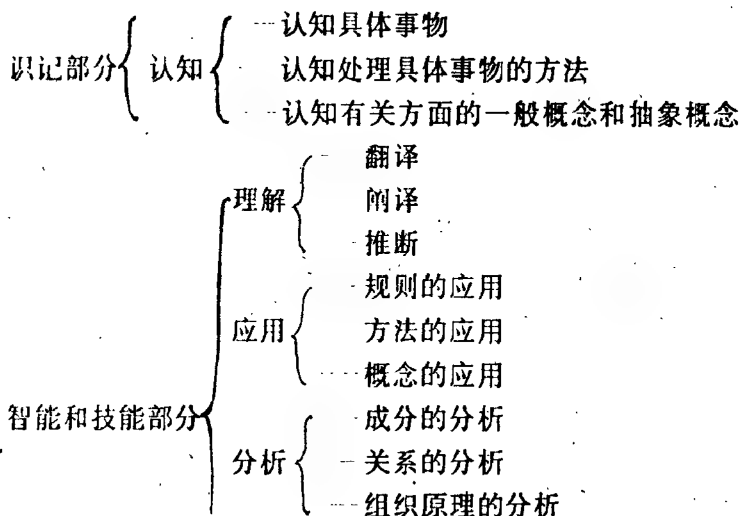
考试在改革，历史观注考试，考试影响未来。美国为首的标准化考试传到中国，引起了许多教育改革者的激动和振奋。他们把这种考试方法在不少杂志上进行介绍和探讨。特别是B·S·布卢姆（美国著名心理学家及课程权威，芝加哥大学教育教授）的认知领域学习目标分类学说用于命题的分类依据。引起了教学工作者的很大兴趣。对我国的考试制度和方法产生了强烈的反响和震动。触发和推动了我国的考试改革，给中国教育的发展予以良好的促进作用。

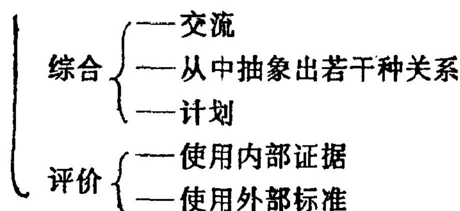
B·S·布卢姆认知领域的学习目标分类学说一直得到

国际上的公认。尽管这种分类还不如化学、生物等学科的分类，但布卢姆的学习分类已证明极其有用，而且已为很多国家的教育家所采纳。严格地说：这种分类是学习目标分类而不是学习本身分类。但它同时为两者提供了十分有价值的分类方法。

布卢姆的学习目标分类，反映了学习结果由简单到复杂、学习水平由低级到高级的层次，对于我们进行教育测量，做到既考知识又考能力，妥善处理两者的关系，并定出能力考查的地位，有很大的启发。不少教师用它来作为科学命题的依据。然而，我们只要回顾一下中国教育的发展，就会察觉到公元七世纪唐代考试方法与二十世纪B'·S·布卢姆的认知目标分类，有一种无懈可击的"巧合"。这真是教育史上的奇迹。

布卢姆学习分类，其中道出了六大类学习，它们分列出了十七小类：





唐代科举考试科目虽多，可其方法不外以下四种（口试）：帖经、墨义、策问、诗赋。口试方法，是对考生当面进行问讯，其内容无明文可查。不知其详。

（一）帖经与“认知”

帖经考试方法是唐代考试一种主要方式。凡“明经”“进士”“明法”“明书”、“明算”各科均须帖经。什么叫帖经？据《通典》记载：“帖经者，以所习经，掩其而端，中间开一行纸为帖，凡帖三字，随时曾损，可否不一。或得四得五为通。”简单地说，就是主考者任择经书中的一页，用两张纸盖着左右两边。中间开一行，另裁纸为帖，帖盖数字，令被试者写读出来。帖的字数可以随时增减，一般要答对四、五条才算及格。帖经这种方法不但科举采用，学校检查学生学习情况时也采用。唐代学校的旬试采取的方式就是帖经。这种考试偏重于记诵，只要求学生记诵能力，而不是十分重视思维能力。布卢姆认知领域里第一层次为“认知”，所谓“认知”实际上也是一种能力，只不过它仅仅是指记忆能力。记忆水平的学习是一种低水平的学习，仅仅有记忆力参加而获得的学习结果，也是一种低水平的学习结果。因此，可以看出：帖经实际上是考知识的记忆。虽权德舆并不以帖经为最好的办法，但认为帖经可保证有基本知识。目前，英

语考试中的clost就是源于帖经这种方法。

（二）墨义与理解、应用

墨义就是笔答有关经义或其生疏的考试。除秀才、进士两种不问大义以外，其它许多科都要问大义。这种考试方式，是为了对帖经这种方式补偏救弊而提出的。公元廿五年的诏书就承认"明经以帖诵为功，罕穷旨趣"。为了矫正这种偏向，才决定增加"口问大义"这种新形式。墨义的方法是主考者出题，由应试者作简单的笔答。如问："子以四权，所谓四权者何？"答"文、行、忠、信"。

宋吕夷简应本州乡试所存墨义试卷为例：

原题：子谓子产有君子文道四焉，所谓四者何？

对：其行已也恭，其事上也敬，其养明也惠；其使民也义；谨对。

题："见有礼于其君者，如孝子之养父母也。"请以下文对。

对：下文曰"见无礼于其君者，如雁？之鸟雀也"。

布卢姆认知领域的学习水平第二、三层，是理解、运用、"理解"指对已答知识的初步领会。具体表现在翻译、改写、解释、引伸等方面。"应用"指在理解的基础上，将知识应用于新的具体情境中去，这是对单项知识理解的深化。该层次的学习结果表现在直接应用已学过的概念、公式、方法、规律等，解决具体的问题。从墨义考试来看。考试的内容也就是理解与运用的问题。唐代学校中的月试就常用墨义这种方法。

---

①宋吕夷简应乡试所存墨义试卷《文献通考》卷三十

（三）策问与分析、综合

策问也是问答，就是按策问临场撰文回答。现抄录元结在唐代宗永泰二年于道州任内问进士五题之一如下：

问：往年天下太平，仕者非累资序，积劳考，二十许年，不离一尉。至于八总区辖，则当名籍甚者得至焉。今商贾贱类，台隶下品，数月之间，大者上污卿监，小者小辱州县。至于廊庙，不无杂人。如专经以求进，主文而望达者，若不困顿于林野，则必凄惶于道路。今日国家行何道，得九流鉴清？作何法，昨侥幸路绝？施何令，使人自知耻？设何教，使贤愚自兮？[①]

但它比帖经、墨义高深，要对现实问题提出建议。大都涉及当时政治、吏治、教化、生产等方面的问题。如两汉时董仲舒的对策就是当时现实的问题。策问对于考查一个人的才能是较好的考试方式，在唐代及第进士中曾出现不少有才干的宰相郡守。都是用策问来筛选的。布卢姆认知领域的学习水平分类：第四、五层是分析、综合。“分析”指在对有关的单项知识理解的基础上，进一步看到它们之间的内在联系，这是对知识的更深人的理解。具体表现在对多项有联系的知识的比较、区分、间接推断、找出联系、列出纲目、归纳等方面。“综合”，指通过对多项知识的应用，组合成一个新的整体，这样学生可以表现出一定的创造性。这一层次的学习结果具体列表现在组合、编辑、设计创作等方面。唐代时策问的要求，既要熟识经史，又要通知时务，既要有鲜明的主张，又要有写作的技巧，不是学问广博的人是难于应

①《元次山集》，卷九。

选的。现在高考中的有些较难的题目要答对应具备的要求，具体上和策问的要求差不多。可见策问考试的生命力的强大。

（四）诗赋与评价

唐初各科考试都以策问为重，直到高宗永隆二年（公元681年）考功员外郎刘思主认为明经更多抄义条，进士惟诵旧策，都没有实才，奏请进士加试杂文二篇（一诗一赋）是为唐代诗赋之始。但这时并不以诗赋为考试的主体。五字太和八年（公元854年）礼部又罢进士议论诗赋，这时诗赋才和策问居于同等地位。后来，实际上进士科的考试偏重于诗赋了，往往帖经不及格的，诗赋好也可以放过。布卢姆的"评价"指在对多项知识理解和综合的基础上，形成某种价值观念，并以它为标准去衡量，估价其它的事物，具体表现在权衡、估价、支持、反移、评论、赏析等方面。从抽象的角度看，由于诗赋在考试中占重要地位，是因为当时唐诗已经成为运行的主体，而唐诗之盛又因进士科的提倡，二者是互相影响的。可以说是时代的产物和时代的需要。"评价"也有同等的意义。从实际讲：诗赋不但能考察思想，而且也能反映一个人的文化修养和文学水平。"评价"也是要达到这一目的。

总上所述，唐代的考试方法和考试内容与B·S布卢姆的认知目标分类来看，从抽象的意义谈，二者是同构的，用"分类"来命题，结合唐代的考试方法是现代考试的雏形。

四、现代考试的特点

现代考试是随着现代教育的产生而发展起来的，它必须适应日新月异的科学技术和教育事业的迅速发展。它从内容

到方法，从组织的管理到具体实施，均与古代考试有相当大的差异。有的地方简直是无法比拟的发展变化。它以自己独特的方法存在着，显示出与古代考试不同的特点。

第一，从考试种类来看，由过去的科举考试发展到多种考试：有学校考试、升学考试、自学考试等。现代考试作为一种社会现象，评价、控制人的知识水平和智力差异的方法与工具，并不象古代考试使用范围那样狭窄。而现代考试可以实施于任何一个社会成员，也为社会成员提供了应试的机会。作为我们日常生活中常碰见的事，人们也常用它来判断自己与他人的知识水平和智力差异。这种情况，封建社会是无法比拟的。

考试的触角和影响，绝不限在教育领域，在当代，凡是需要选才的部门和单位，几乎都离不开考试。需要某种技能的职业（如汽车司机、飞行员），需要某种特殊才能的职业如演员、时装模特儿，需要较高知识和能力的职业（如医生）、常举行招聘或策定性质的考试，考试真是无时不有，无处不在。一九八三——一九八四年度，美国普林斯顿考试服务处（ETS）组织了一百多个项目的考试。有一百五十多个国家和地区的八百多万人参加了考试，预算开支达一亿四千五百万美元。

第二，在考试内容方面，由主要考核空洞无用的经典改变为主要考核现代科学知识，由只重考记忆和背诵发展为侧重考理解和运用的能力上来。

随着科学技术的进步及其在生产中的广泛应用，科学技术对生产发展的作用越来越大，日益成为推动生产力发展的主导力量。与此同时，教育的生产职能亦即再生产劳

动的职能也日益重要了。八十年代，知识不断"激增"，知识也在迅速陈旧，要学要懂的东西越来越使人眼花缭乱，目不暇接。"未来的文盲不再是不识字的人，而是没有学会学习的人[①]"。如果说古代教育的生产职能主要由父传子，师带徒来实现，学校教育主要是培养统治者的接班人的话，那么现代教育的生产职能则主要靠学校教育来实现。学校教育就同时具有培养统治者接班人和再生产职能与现代的生产资料结合的劳动力这两种主要职能了。这时，传播现代的科学技术，就成为学校教育的主要任务之一了，相应的，现代考试的主要内容，也就由古代的圣贤说教改变为现代的科学知识了，而科学技术的价值仅在于应用，应用它去认识世界和改造世界，而理解它又是应用它的前提，因此，现代考试侧重考对知识记忆的同时，要考对知识的理解和应用。

第三，在考试的组织管理方面，更加科学、更加严密。考前阶段的考试计划的编制，试卷的编、印、送的组织管理，到考场的编排、监考、试卷的装订，特别是评卷的组织管理工作等等，都与古代考试有了重大的变化。

第四、在考试方法方面，既继承传统的考试方法的精华部分，又创造了许多现代考试内容和技术手段相适应的新方法。

为适应考核现代科学技术的需要，人们在实践中又创造了变化繁多的现代计算题，工程设计题科学实验题，并常常把口试、笔试、操作结合起来运用于一项考试。特别是，采

---

①《第三次浪潮》〔美〕，阿尔温．托夫勒，生活、读书、新知三联书店

用机器阅卷给分的多项复杂选择题，是非判断题，甚至还能编制标准化测验，使试题能够一用再用。由于这种考法能够使用现代化的考试手段于设计和编制试卷，分析和修改试卷，阅卷和整理分数，因此，有时人们也专称这种考法为现代考法或"标准化考试"。

随着考试内容，方法的变化和考试的规模，应用的扩大，从们逐步明白了掌握了考试的内在规律，并将某些现代的科学技术应用于考试工作中，使考试发生了质的变化，即由传统的经验考试转变为现代的科学考试。

第三节　考试学的任务和内容体系

一、考试学的任务

考试学是以现代化的科学技术和理论为工具，以考试为研究对象的理论与实践相结合的一门新兴的科学。它的任务是揭示考试的客观规律，探讨科学地，准确地考查人的知识、能力水平和智力差异的方法，从而指导考试实践。

现代考试，是面向全人类的考试，其规模是古代考试所无法比拟的。科举考试的科试、分试，诚然也是在各地同时举行的面向社会的考试。但各地的考试是分别出题，分别组织的，实际是许多个小规模考试的组合。现代的考试是全国或地区规模的考试，有时是全球性的。如象美国普林斯顿考试服务处，就是面向全球的。因此，考试是一项十分复杂的组织工作，研究大规模考试的科学管理办法是时代的需要，也是考试学的主要的任务。

随着社会的发展，信息的作用和价值越来越大，考试作为获得信息的手段，其应用也越来越广泛。它不仅用于教育领域，对于如象职业能力鉴定，智力诊断和能力发展预测等，都起到巨大的作用，不同类别的考试，其组织实施过程，要求获得准确可靠的信息，以及使用不同的方法处理信息等的不同特点和特征。这是古代考试不曾出现的。研究不同考试的不同特点和处理信息的不同方法，是社会的需要，也是考试学的重要任务。

考试是让考生在规定的时间内，按指定的方式，解答事先根据考试目的编制的题目，对解答的结果评等判分，为主考者提供考生的知识能力状况和智力差异的信息。为了使获得的信息准确可靠，编制的试题的内容应有足够的代表性，对被试者的特点应有足够的针对性，并尽量减少试题解答和评卷中无关因素的干扰。而就必须研究、掌握试题取择（使考试内容有足够的代表性）和编制（使对被试者有足够的针对性）的科学方法。必须研究、掌握试题解答（使被试者正常发挥水平）和评阅（使之客观、公正）过程的科学控制方法。研究试题取择和考试实施的科学方法，变经验考试为科学考试，是考试学的最基本又最重要的任务。

二、研究考试学的意义

我们研究考试学，就是为了探索、认识、掌握考试的规律，用以指导考试工作，使考试在育人，选人和用人工作中更科学地发挥作用。既有十分紧迫的现实意义，又具有深远的历史意义。

考试曾发端于我国西周时代，盛行于我国。比起欧美国

家，要早两千多年，但直到今天的考试，还是那么不完善和较为落后的。研究工作和研究水平远远落后于发达国家。由于我们长期不重视考试的科学研究，致使在许多国家已经进入考试标准化，考试手段现代化的今天，我们仍然固守传统的经验考试办法。虽有局部地区采用现代考试手段，但在大规模范围内，面对高等学校招生考试，高等教育自学考试等千百万人参加的大规模考试，我们还在搞“人海战术”“手工操作”，由于我们的考试属经验考试，在考试时缺乏可靠性、有效性。对考试分数研究不多，迷信分数，几分甚至一分之差就可能改变一个人接受教育的机会和工作方向。因此，研究考试科学、建立考试学，是实际工作向我们提出一项十分紧迫的任务。

在世界范围内兴起的新的技术革命，对于人才的培养和选拔工作提出了新的更高的要求，也给考试的研究工作提出了新的课题。

历史的发展要求采用技术发展的新成果，武装和改进考试工作，使考试工作手段和方法不断地现代化。考试工作中广泛使用电子计算机，特别是考试工作的核心环节——命题工作的电脑化，将成为必然的趋势，由此将带动整个考试工作的现代化，考试内容现代化，命题方法现代化，考试组织管理现代化。

考试与教学的分离，是教育评价的一大进步，在新的条件下，作为教育评价的考试与教学的融合。很可能是教育评价的又一次大的进步。近年来，在某些发达国家中又出现了采用机器分散教学的趋势，在那里，考试已同教学融为一体，考试与教学过程融为一体，成为学生学习知识，提高能力和

随时取得反馈信息的一种方法，从而改变学校考试旨在区分学生优劣的传统做法。为此，研究新技术革命对考试工作的新要求，使考试工作在初步科学化的基础上进一步现代化，是考试学研究的重大课题。

三、研究考试学的根据

新时代、新的社会现象，需要考试要科学、需要加强对考试的科学研究，建立为之服务的考试学。那么研究和建立考试学的根据何在呢?

首先，我国古代考试的丰富经验为我们研究、建立考试学提供了重要资料。学校考试制度在我国已有两千多年的历史。对我国教育、政治、经济等方面有重大影响。科举考试制度建于隋唐，距今有一千三百多年的历史。祖先为我们在考试学方面积累了丰富的有巨大研究价值的考试资料，是一份难得的遗产。遗憾的是，我们对这笔遗产的发掘、利用得很不好，虽然评价科举考试制度的得失的文章在报刊上偶有所见，但至今尚无一部系统研究古代考试资料的书籍。现代考试是古代考试的发展，无疑对古代考试之优秀部分应予继承和发扬光大。发掘、整理和批判地研究古代考试的丰富资料，对于研究、建立考试学，指导今天的考试工作，具有重要的意义。

其次，建国以来的考试工作的实践，成功的经验与失败的教训，为我们研究、建立考试学提供了坚实的实践基础。我们要根据考试工作提出的实际问题确定考试学的研究题目，各级各类学校的考试实践，高等学校招生考试的实践，高等教育和中等教育自学考试及其他考试的实践，为考试研究提

出了许多课题和丰富的感性材料。这是研究考试的实践基础，不离开考试的实际需要来研究考试．几年来许多从事考试工作和考试研究工作的同志，从我国考试的实际出发，在考试内容、方法、方式和制度等方面，进行了卓有成效的研究工作。陆续发表了一批有影响的研究文章。我们要充分利用这些丰富材料，来认识考试的客观规律，建立考试学。

再次，外国考试的成功经验和考试学研究的成果，对于我国研究和建立考试学具有重要的借鉴意义。我们应该系统研究，虚心学习外国考试工作的先进经验和考试研究的积极成果。从我国的实际情况出发，吸收其科学的合理成份，特别是在本世纪二、三十年代以来，西方许多国家在考试工作和考试方面的成果、如像客观性考试，标准化考试，智力测验。某些国家还设有研究和组织考试的专门机构等等，来推动我国考试研究和考试工作的发展。

第四，教育心理学、思维科学、人才学和统计数学、模糊数学等邻近学科的开创和发展，为考试学的研究提供了极为有利的条件。考试学是研究如何科学地考査、测量人的知识和能力水平的学科，它与研究人的知识和能力习得规律的教育心理学，与研究人才的智能结构和成长规律的思维科学和人才学，在知识体系上有着密切的内在联系，它在研究过程中，必然要应用教育心理、思维科学和人才学的研究成果。考试学的研究常常要运用统计分析的方法，进行定量分析，统计数学为这种分析提供了方法的基础。由于事物的特性精确与模糊导致考査、测量人的知识、能力和智力差异的模糊性，常用模糊数学的方法来解决定量定性的研究。计算机的广泛应用促进了统计数学，模糊数学在理论和方法上的新进

展，这是研究考试学的极为有利的条件。

四、考试学的基本体系和主要内容

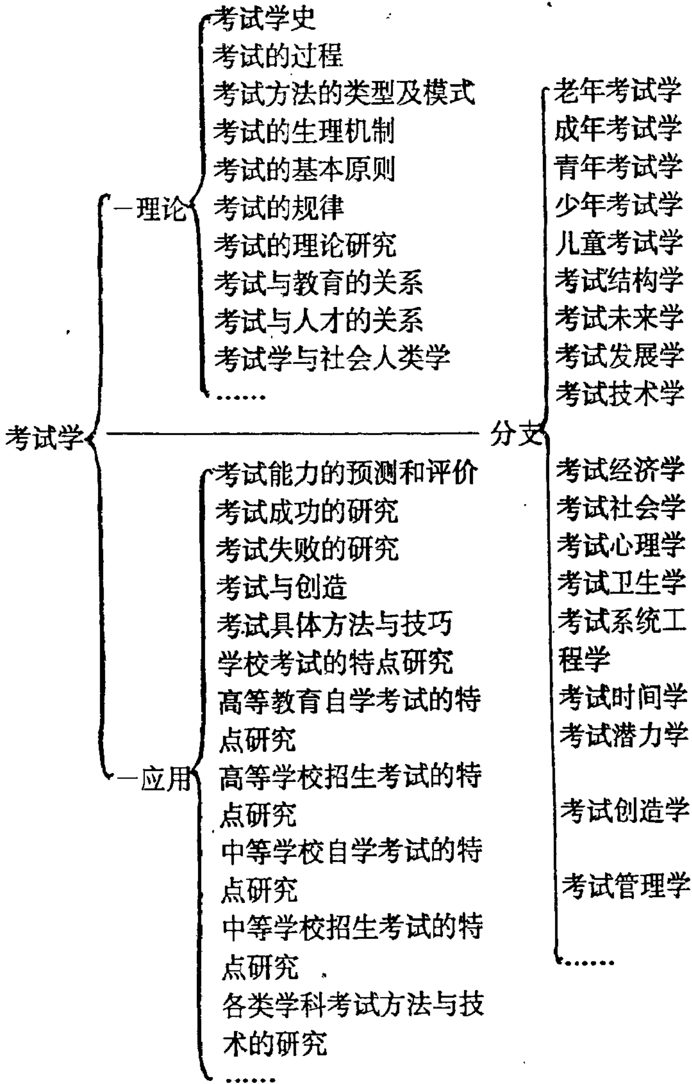
考试学是一门研究考试理论与指导考试工作实践的学科，它应该系统地回答各类考试实践中提出的问题，并形成自己的科学体系。我们认为考试学的体系结构应包括理论、应用和分支三个部分（见表一）。

由于考试学还是一门正在形成中的新的学科，它的体系和内容尚在建立和探索之中，因此，初建的考试学，应当把论述的重点放在一般考试实施过程的科学分析上和对于我国最重要的几种考试制度专题分析与研究上，这样一方面，它作为一门应用学科，对于当前考试工作才有更强的针对性和指导作用，以满足实际工作的急切需要；另一方面，它作为我们对考试工作理论探讨的第一步，又为我们在实践中对考试进行更高程度的理论概括，更深层次的科学分析和分支学科的建立奠定必要的基础。据此，本书包括以下几个方面的内容：关于考试的功能和作用；组织考试工作的基本要求和考试质量指标的理论分析；关于一般考试实施过程的科学分析；关于管理考试科学现代化的理论探讨与展望。

第四节　考试学的基本特征

考试学作为以研究考试的科学，它不仅有自己特定的研究对象，而且也有自己的基本特征。

表（一）


考试学
理论
考试学史
考试的过程
考试方法的类型及模式
考试的生理机制
考试的基本原则
考试的规律
考试的理论研究
考试与教育的关系
考试与人才的关系
考试学与社会人类学
……
应用
考试能力的预测和评价
考试成功的研究
考试失败的研究
考试与创造
考试具体方法与技巧
学校考试的特点研究
高等教育自学考试的特点研究
高等学校招生考试的特点研究
中等学校自学考试的特点研究
中等学校招生考试的特点研究
各类学科考试方法与技术的研究
……
分支
老年考试学
成年考试学
青年考试学
少年考试学
儿童考试学
考试结构学
考试未来学
考试发展学
考试技术学
考试经济学
考试社会学
考试心理学
考试卫生学
考试系统工程学
考试时间学
考试潜力学
考试创造学
考试管理学
……

一、实践性

实践性是考试学立足的基点。考试学是一个国家、一个地区的经济、科技、社会实践和教育实践的产物，考试的形成和发展的实践经验是考试学的科学基础。实践是检验真理的唯一标准，检验和判断考试合理性和科学性的唯一办法，就是科学地分析实践效果。即或学习和借鉴他人的经验，也要在实践中试验和消化才能变成自己的东西。真正的理论也要在实践的基础上，把感性认识上升到理性的认识，把零散的经验上升到理论原则，理论又回过来指导实践。如此不断循环和发展，是考试学立足的基点和成长的生命线。考试学所侧重的考试，既来源于经济、科学、教育等的社会实践，又适于经济、科学和教育的社会实践。它是一门实践性很强的科学，具有鲜明的针对性、现实性和可行性。因此，实践性是考试学的重要特点。

二、科学性

任何一门学科都有自身的规律及其应遵循的科学原则，考试学也不例外。研究建立和发展考试学，要遵循以下几个基本规律。

（一）按照考试规律研究考试

考试既要研究与人的发展关系，又要研究与社会发展的关系。是属特殊社会现象。它可以借鉴其它社会科学和自然科学的理论与方法，但不能照抄照搬。我们研究考试学的根本目的，是为了探讨高效而合理的考试途径，改革与完善既符合考试规律、又可以多出人才、快出人才、出好人才为目

的的考试体系。因此，考试的客观规律是考试必须遵循的一条基本规律。

（二）运用学科的理论和方法研究考试

随着科学技术的迅速发展，在科学研究过程中自然科学与社会科学众多学科的概念、原则和方法，正在相互渗透，相互吸收，那些对自然科学或社会科学传统的单科研究方法已经不能适应现代科学研究的需要。德国物理学家普朗克曾经说过："科学是内在的统一体，它被分解为单独的部门不是由于事物的本质，而是人类认识能力的局限性。实际上，存在着从物理到化学，通过生物学和人类学到社会科学的连续链条，这是一个任何一处都不能被打断的链条……"。考试学是由众多学科理论和方法而形成的新学科，它必然烙上这些学科的特征及其方法的印记，同它们有一定的亲缘关系和相互借用的地方。这既是考试学形成和研究的特点，也是现代科学发展和研究的特点。

三、整体性

整体性是考试学的基本出发点。考试是一种十分复杂的社会现象，它同整个社会各个方面都有关系。因此，考试学的一个根本任务，就是对一个国家、一个地区的考试实施进行总体设计，进行智力资源的全面开发和利用。不论是考试的宏观结构，还是微观结构，都要明确考试在社会中的地位和作用以及与其它社会现象的关系。因此，考试又一重要特征，就是要立足于整体，科学地协调各部分的关系，以达整体的优化。

四、动态性

事物的稳定性是相对的，发展变化才是绝对的，考试也不例外。研究考试的目的之一，是要建立一个最佳的合理的考试结构。而考试结构的合理性只相对稳定于一定的历史阶段。今天是合理的，但当经济和科学技术发展到另一个水平时，就自然会有可能不合理的因素了，就要改变之。考试的整体结构，不仅其内部诸要因及其相互关系，也是不断变化和发展，而且同其它有关系统之间的相互关系，也是不断变化和发展的。这种变化是经常发生的。其中任何一部分、任何一层次的变化都可能影响整个考试同外部环境的关系。这种发展和变化的根本目的，在于系统内部诸要素之间和系统外部诸关系的动态相关性。相关的性质又与组合的形式有关。在任何时候都不能用静止的观点，分离的观点、单因素的观点来看考试，防止不合理的组合，建立合理组合。考试学应当借鉴和运用现代管理科学提出的"动态相关性原理""时空变化性原理"和"信息传递性原理"来研究考试这个动态综合体。

五、超前性

考试是一个指挥棒，要更好地应用考试来服务于社会，就必须研究它的发展。确定出它的方向，掌握住考试演变，社会的发展迫使研究未来的考试，只有加强未来的考试研究才能适应社会发展对考试学提出的新要求。因此，考试学有一显著的特征就是超前性。

第五节　考试学的研究方法

任何事情都要讲究方法，任何学科也都要建立本学科的研究方法。考试学的研究方法，就是人们运用自己的智慧总结考试经验，认识考试规律，提出考试理论，探索考试的手段，寻找观念世界与现象世界的联系，寻找与外部诸关系的联系，运用科学实践和理论思维技巧，建立考试实践和理论思维技巧，建立考试学体系，改进考试的科学研究工作。

马克思主义哲学是研究和建立考试学的理论基础和方法指南。它是关于自然，社会和人类思维运动最一般规律的科学，它为各门具体科学的研究，提供了科学认识论和方法论的基础。我们研究考试学，应当自觉地坚持和运用马克思主义的认识论和方法论。

由于现代科学的迅速发展，科学的高度分化和综合，使科学知识更为复杂，更为多层次和多序列，几乎每一门科学都形成了宏观和微观层次。实用部分和基础科学部分，纵向联系和横向联系的结合，以科学技术群落的形式出现在学科领域中，进行着边缘学科的研究和综合性的研究。传统方法论手段面临新的变革，“方法论手段的发展将保证从实物水平和研究转到系统水平的研究，从单义研究转到多义研究，从线性研究转到非线性研究，从一种单位的研究转　到多种单位的研究[①]”。这种科学研究方法论手段的转变趋势，必然影响考试学的研究方法。客观事物是错综复杂的，一般都

---

①〔苏〕B．П库兹明，《马克思理论和方法论中系统论原则》第159页

需要几种研究方法的配合，才能揭露事物的本质，寻出事物发展变化的规律。而且，常常是在同一研究项目的不同研究阶段上，也要运用不同的研究方法。下面介绍和讨论几种有关的研究方法。

一、经验方法

为了研究理论，探索规律，运用历史的方法，调查的方法，实验的方法，观察的方法，比较的方法对考试学的形成、沿革、现状和他人的实践经验进行总结是十分必要的，是不可忽视的研究方法。

历史法是运用文献史料进行研究的方法，运用分析过去的考试的实践和理论来认识考试的历史和现状及其演变的过程。继承前人的经验和成就，吸收前人的经验教训，使“古为今用”。观察法，调查法和比较法都是从研究国外考试的现象出发，从过去和现在的实践中搜集需要的事实材料，来总结经验，研究理论，探索考试规律。

实验法，也是科学研究的经验方法。它和其它经验方法的区别，在于对研究现象进行一定的人工控制，以便较为准确的确定事物的矛盾，探索现象的因果联系，检查预想方案的效果。验证设想的可靠性。按照建立考试学的原则，方法和设想方案，选择适当的地区和学校进行单项或总体结构改革试验，是研究考试学实践与理论的不可缺少的重要方法。也是当前进行考试改革特别值得重视的方法。

二、理论方法

科学研究的理论方法，是以某些既定的前提，运用“分

析和综合的结合”对科学理论的概念、定义、范畴、规律和理论体系进行各种表述及其认识活动的方法。这类方法，是以某种程度的间接性为特征的、是考试学的重要研究方法。

三、思维方法

在科学研究的过程中，运用思维科学在表象、概念的基础上，通过判断、推理、想象、创造等认识活动，进行“判断、推理、想象、创造”适合我国国情的社会主义考试体系的科学研究工作。前述的经验的方法、理论的方法也都要经过科学的抽象思维才能完成科学研究的任务。

思维方法，根据我国著名科学家钱学森教授关于思维学的立论，又划分为逻辑思维方法、形象思维方法和灵感思维方法三个亚类，包含着若干种研究方法。例如：

（一）类比方法

类比方法亦称类比推理方法或类推方法。一般分为定性类比方法和定量类比方法。它既要借助于原有知识，又不受原有知识过分的约束。不同事物联系起来，异中求同，同中见异，产生新的知识。同类事物有相似性，适于类比法。从广义的哲学观点来看，一切事物之间都存在某种程度的相似性，因此，类比方法不仅用于同类事物之间，也可以用于不同事物之间。现代科学技术的变化、综合、相互渗透，为类比方法的应用提供了广阔的活动场所。在产品设计和仿生学中，类比方法都发挥了显著的作用。尤其在某个科学领域的开拓时期，在一个领域向另一个领域过渡的时候，在学科临界产生边缘科学的时候，类比方法都明显地起着启示、探索、开路、创新的作用。在考试学的开拓时期，把类比方法和演

绎方法、归纳方法以及其它方法结合在一起，综合运用，必将有力地推动考试学的发展和完善。

（二）移植方法

将一个学科领域中已经发现的理论知识和行之有效的研究方法，移植到其它学科领域中去的方法称为移植方法，它具有综合方法和类比方法的特点。科学发展的历史表明，多数的新理论、新技术或新方法，都可不同程度的应用于其它领域之中，新的理论和方法一经出现，聪明的科学家即从各个可能的角度予以观察和研究，并将它与有关的知识联系起来，寻找科学研究的新途径、新发现。移植方法告诉我们，运用几个学科的理论知识和研究方法，去研究考试，完全可能开辟一条创新学科——考试学的途径和方法。维纳创立的控制论，就是借助了多学科的移植和综合作用，从而产生了边缘学科这个新的学科门类，这对我们的启示是很有益的。

四、系统方法

系统方法是人类思维和现代科学技术发展的结果，它具有高度的抽象性和高度的综合性，是以综合为基础的。然而，它不是不要分析，而是在综合过程中把分析有机地结合起来。它从综合出发，在综合的基础上进行分析，再回到综合，它的工作公式不是“分析——综合”而是“综合——分析——综合”。运用这个程序来研究考试系统。其一，要把它作综合体来考察，了解这个系统的各个组成部分（元素）之间的关系；了解这个系统的结构，把各个组成部分（元素）以他们之间的相互关系为网，联结在一起，构成考试系统结构模型，

这就是第一步综合的结果。其二，在综合基础上进行分析，把考试结构模型分解为几个互有作用关系的子系统，这些子系统又可分解为下一层次的子系统，直至内部关系比较简单不需要再分解的子系统为止，逐层分析考察每一个子系统。这样的分析是按照事物相互作用关系在整体结构中进行的分解的。分解后显现互有联系的层次结构。每一次分析的结果要反馈到上一层去，与整体的要求进行比较，按照比较的差距重新修改整体结构模型。根据反复分析与协调的结果来确定系统的构成方式和活动方式，进行系统的合理组合设计，实现部分与整体的统一。

根据系统工程方法的要求，在步骤上按照统一规划，建立模型和仿真，决策和实施去进行。

统一规划，首先要摆好问题，确定目标，要详尽地占有材料。通过调查研究，尽量地全面收集考试体系的历史、现状及发展趋势的资料和数据，按照考试任务的要求提出要达到的目标。

设立模型与仿真，就是要依据统一规划的目标要求，把整个考试系统的信息流程、物质流程和人才流程复杂关系定量或定性地表达出来、并画出结构网络图。

决策和实施，就是把上面得到的结果，进行分析比较，选择最佳方案进行实施。

五、数学方法

马克思说过：一种科学只有在成功地应用数学时，才算达到了真正完美的地步，才算真正发展了。数学方法为科学研究提供清晰精确的形式化语言、推理工具、抽象能力、数

量分析和度算方法。它具有横向移植的特征，可以横着伸向一切学科领域。当然，考试学也不例外，它同各种研究方法相互结合，进一步揭示考试体系结构的本质和发展规律。

考试学是一门新兴学科，在这个领域内应用数学方法同自然科学诸学科具有不同的特点，由于考试系统是一个多因素、多变量、难测度、有较大信息量通过的多科的动态系统，在运用数学方法时，必须借助于“载体”才能把数学方法用于考试的研究中，可以充任“载体”的科学方法有实验方法、技术方法、系统方法和控制论方法等等。所以，在研究考试时，要首先借助于系统方法，控制论方法分析和确定其构成，提出数学模型，才能在计算上进行模拟。随着科学研究的深入和科学技术迅猛的发展，越来越依赖于数学方法获得新的科学研究成果。

在介绍和讨论考试学的研究方法时，我们应注意吸收现代新科学和新方法，这对于形成考试学自己的完整理论体系是有益的。也是十分必要的。也只有吸收有关学科的理论和方法，才能对考试学的研究更加深入，也才能尽快地、科学地建立考试学的理论体系和方法体系。

第二章　考试的作用、测量特点、功能与设计

第一节　考试的作用

一、考试的目的

考试的目的，各家的说法不一，但有一点应该是肯定的，那就是它会随应用的场合不同而不同。我国封建社会为"取士"而开的考试，早于中世纪西欧各国的就业资格考试，与学校教学成绩的考试所欲达到的目标当然不会相同。从教育方面讲，历史上不少国家都曾确定过以提高教育质量和学习水平为考试的宗旨。为维护教育水准，他们采取过严格的入学、升级和毕业考试办法。在现代西方的教育文献中，有人主张从理论上讲，考试的目的应有两个："一是测量受考试者过去的成绩，一是评价其未来的才能[①]"。当然，如果从一般测量学的角度看，这种简单的说法似乎无可非议，但是，如果从教育学的观点衡量，就不十分妥当。因为它并不能全面概括教育和教学方面的要求。在教育界，对此也常是各持所见。有人认为"教学成绩考核的目的在于通过考前的复习和考后的补习，系统地总结和巩固已学过的知识。也有人说

① 参见《学会生存》1979年版第118页。

是为了“评定教学质量，确定升级留级。”在有的实行学分制的大学，把它说成是“为了取得学分”等等。

实践证明，如何确定考试的目的，关系到怎样对待考试，用何种方式方法考试，以及考试后的效果如何等一系列的问题。因此，明确考试的目的，往往成为搞好考试的前提。就考试的发展和现状，其主要目的归结为下面几个方面：

一、根据传统教育观点，考试是为了对学生的知识和技能进行阶段性的和总结性的检查与评定。考查学生按照课程大纲规定掌握知识的数量和质量。技能的准确与熟练程度，以及运用所学知识与技能分析解决实际问题的能力的发展水平，据此评定学习成绩，并使之成为升留级、取得学分、毕业的主要依据。

二、从现代教育观点来看，考试是促进学生智能发展的环节之一，考试的过程（包括组织、实施和总结的各个阶段）是平时教与学的两个方面的深化和提高。在考试环节中，学生在教师指导下，对所学知识和技能加以归纳整理，使之系统化、条理化，进一步领会所学知识之间的内在联系和规律性，力求达到融会贯通，加深理解。这就是培养和发展学生思维能力、创造精神、增强自学能力的过程。

三、教育心理学认为，考试是激励学生学习兴趣和进取精神的有效手段之一。考试可以帮助学生自我认识。了解自己的学习状态，认识自己在学习目的、学习态度、学习方法、思维方法、意志品质以及体质状况等方面的长短。同时由于恰当的考试方法可以使学生发现和体验自己学到的知识、能力对于认识、改造客观世界的能动作用。这样，他们的学习兴趣和学习动机都得到发展和强化。在教学过程中达到道和

业的结合，志和趣的统一，不仅有了源于高尚理想和远大目标的远景性动机，而且有了远景性动机背景上出现的近景性的具体动机，则他们的兴趣更深刻，进取精神就更强烈。

四、从现代管理学的原理来看，考试在教学过程中的地位，即“P—D—C—A”循环中的后两个环节——检查和总结，通过考试可以检验教学效果，获得的信息可为改进教学工作提供指导。从系统控制的观点看，考试获得的反馈信息，经分析处理，及时地进行调节，保证最大程度地逼近质量目标。这样，可以使教师了解学生的学习现状和发展趋势，也了解自己，总结教学经验，更好地把握教学规律，改进教学方法，也可以使教学管理系统对各要素的工作质量进行分析评价。对系统状态的不均衡因素加以调节控制，从全局上调整、修改、充实和深化原来教学管理的决策和措施。

五、考试受一定社会的政治、经济的制约，又在社会上产生广泛的影响。一种考试方法或考试制度在学校中和社会上所产生的影响会远远超出考试工作本身。超出考试者和被试者的范围。这就是我们现在常说的考试的“指挥棒”作用。产生这种影响的主要原因是：考试是以选拔人才、识别和发现人才、正确使用人才为目的。这就是考试的社会目的。

实现上述目的，就体现了考试应起的作用。

二、考试的作用

关于考试的作用都是众说纷纭，有的说：考试是检查教学效果的主要手段；有的说：考试是对所学课程复习、巩固、提高的重要环节；有的说：考试不仅是考查学生的学习质量，更重要的是考核教师的教学质量；……这些说法虽有一定的

道理，但也有片面性，不能完全准确地反映问题的实质，那么考试的作用，究竟是什么呢？概括为反馈作用和鉴定、选拔作用比较恰当。

从科学管理的角度来看，考试是一种信息反馈。完善的成功的考试 不仅可以比较全面，比较准确地反映出学生所掌握知识、能力水平和智力差异的状况，而且可以反映出教师所选择的教学内容是否恰当。采用的教学方法的成败得失。教师和管理者利用得到的这些信息，就可以不断地调整和改进教学工作。反之，不严肃、不认真的考试，很可能提供失真的甚至是虚假的信息，根据这种信息反馈进行决策，必然要造成失误。因此，改进考试工作，提高考试质量的目标之一，就是要提高信息反馈的质量。

总上所述，考试是学校检查教学效果、取得反馈信息、评定学生的学业成绩，衡量学生是否达到规定的数学目标的重要方法，也是教育行政部门检查各学校教学质量的重要手段。

考试的鉴定和选拔作用，指通过考试对应试者的知识、能力水平的发展和智力差异等诸方面作出比较全面、比较正确的评价。作为鉴定和选拔人才的依据。随着社会的发展，社会分工越来越细，许多工作岗位虽然并不一定需要很高层次的知识分子，但它对专业知识和技能的要求也逐步提高，因而也时常需要采用考试的办法对应招人员进行选拔。可以预见，随着我国现代化建设事业的发展，需要进行选拔考试的岗位将会越来越多，考试作为选拔手段的应用将更加广泛。高等学校招生的考试，学校中根据考试的成绩对学生进行升留级的处理，就是考试的鉴定和选拔作用的体现。鉴定选拔

作用是考试十分重要的功能，它不仅直接关系着评价和选拔人才工作的质量，而且鉴定、选拔工作的准确性对反馈信息的可靠性密切相关。人才具有不同的类型，不同的层次，不同的水平，不同的特点，他们在知识上有不同的水平，在能力上有不同的倾向，在智力上有不同的差异，为了量才使用，使之各得其所在，显然是十分必要的。考试是使之成为这种分辨和识别可行的方法之一。特别是在鼓励人才合理流动的时候，人才流向的单位难于采用实践考查法考核流入人才的长处和特点，考试就成为经常采用的方法。

三、考试的局限

任何手段都有其优和劣，考试也同样如此，它有一定的适用限度和局限性。也有其短处和方向性。

考试的局限性主要表现在以下几个方面：

第一，考试带有主观性。

任何考试都要求尽量地客观，即要求同一考试在由不同的人实施时所得的测量值尽可能一致，但是就考试内容，考试标准，考试对象，由不同的人主考等等，都由考试的目的所决定的。目的是由于社会的需要决定，因此，考试就带有一定程度的主观性。就目前考试的发展状况，有些考试设计和测试实施都是凭借主考人的个人经验进行的。而不同的主考人，对考试目标的理解，试题和试卷的编制，正确答案的认识，给分标准的掌握，又总是不可能完全一致的。为此，研究考试的客观性与主观性之间的矛盾始终是考试研究的重大课题。

第二，考试的偶然性和分数的局限性。

考试有偶然性，特别是一次性的考试，这里的偶然性就是说学生的成绩的的不稳定性，即时高时低。

试题内容是从总体抽样出来反映总体的内容。考试在某程度上是抽样测量，抽样测量常带来的统计偶然性。就某一次考试来说，它的试题值就可能与总体的真值有较大的偏离，不可能无限度地增大试题的样本和考试的次数，使考试的平均值趋近于总体的真值。如果选取的试题恰巧是某人没准备好的内容，他的考试成绩就会低于他的真实成绩，如选取的试题恰巧是某人准备好的内容，他的考试成绩就高于他的真实成绩。

考生在应考时的临场发挥是否正常，与考试的成绩（等级或分数）是正相关的。考生是在规定的场所、限定的时间内、有监督的情况下，按指定的方式回答问题的。在这种特殊的气氛和种种约束下，有些人常常失去了平时具有的思维敏捷性和灵活性。有时甚至会出现把平时牢记在心的东西都遗忘的怯场现象。而有些人有时突然可能出现最佳的思维状态，思潮澎湃，涌向口头或凝聚笔端，从而使临场发挥呈现不平衡的状态。

考试的偶然性给考试成绩带来了局限性。因此，我们不要把考试的成绩绝对化，企图用考试的分数说明一切问题，也不要把考试绝对化，滥用考试。比如选拔人才，就有推荐法和考试法两类办法。应尽量避免仅根据一次考试的结果，决定人的升降取舍。

第二节　考试与测量

一、测量的概念

测量（measurement），也称度量，本质上是一种比较的活动，是通过将被测物体与参照物体进行比较，从而对被测物体赋值的过程。美国物理学家默根（M·Mesken）在谈到物体物理量测量时指出："度量就是对某一物理量与一选定的标准进行比较。例如，想要决定一张桌子的长度，就要选定一合适的参照标准，并把它一次接一次地摆下去，直到比较完毕，度量的结果是以桌子等长的参照单位的数目来表示。这种决定'多少'的过程叫度量[①]。"这就指出了测量是一种与标准物相比较，以对被测物体的某一属性物理定量进行赋值的过程。

测量本质上是通过比较而对被测客体赋值。在测量中比较一般有两种形式：一是直接比较。为了确定两物体的长短、强弱、好坏等等。人们将两物体直接放在一起，加以比较，以获得对象某一属性相对强度的信息，这种比较叫直接比较。直接比较是一种简单易行的方法，但是，它的应用受到许多限制，比如，对在时间和空间上有一定间隔的对象，运用这一方法，可能就会有比较大的困难；比较的另一种形式是间接比较，间接比较是以标准物作为中介的比较。比如：为了测定两物体的相对长度，我们可以用米尺作为比较的标准物。

① ［美］M．默根著、暴永宁等译：《物理科学及其现代应用》，科学出版社1983年版，第79页。

100米是50米的两倍，所以物体所用的长度是物体的两倍。这样，我们可以不实际地把两物体放在一起，而通过米和厘米这样的中介进行比较。就其适用范围的广泛性来说，显然，间接比较具有更多的优越性，它不受时间间隔和空间距离的限制，因此，在自然与社会现象的测量中间接比较是一种常用的方法。但是，也应看到，间接比较是以标准物为中介，为尺度的比较。比较的结果以选定的标准物为转移，因此这类比较对标准物有较高的要求，它们必须有极大的通用性，为大家一致认可。比如，在物体物理几何性质的测量中，这些作为中介的标准物就是克、厘米等众所周知的物理单位。这些单位是为大家所接受的。如果这些单位没有通用性，那么测量的结果就很难为大家所接受和承认。美国学者荷尔顿（G·Holton）说："大体说来，几乎每项在实验室作出的测量，在某种意义上都是可观测事物和他们的全体同行所接受的若干闻名的标准二者之间的比较。科学中许多成功和飞速发展都依赖于这些简单的真理。因为显然只有在这样的情况下，许多人员的精力才不致于经常耗费在对意义和程序规则进行毫无结果的争辩上[①]。"事实上，在教育科学的研究中，人们常常进行着许多毫无结果的争辩，缺乏这种约定的标准就是一重要原因。

直接比较与间接比较这两种方法在教育测量中都有着重要的运用。关于这两种方法在教育测量中各自的特点在本章的后半部分，我们再作适当的探讨。这里需要提出的是，直接比较与间接比较是根据作为比较过程标准物的特点而加以

① ［美］G．荷尔顿著，张大卫译，《物理科学的概念与理论导论》，人民教学出版社1982年，上卷，第254页。

区分的。它同根据物体的被测的属性与人们所要想测的属性直接相关程度来区分的直接测量的概念并不相同。

二、人的知识、智力、技能的可测性

知识最早的意思指相知。相知，指熟悉的人。《管子·入国》："不能自生者，属之其乡党知识故人。"《吕氏春秋·慎人》："人有大臭者，其亲戚兄弟妻妾知识，无能与居者。"亦指亲友。韩愈《赠别元十八协律》："知识久去眼，吾行其既远。"随着历史的发展，社会的更替，知识的本意也发生了变化。它是人们在社会实践中积累起来的经验及其概括总结，从本质上说，知识属于认识的范畴。毛泽东指出："什么是知识？自从有阶级的社会存在以来，世界上的知识只有两门，一门叫做生产斗争知识，一门叫做阶级斗争知识。自然科学、社会科学，就是这两门知识的结晶。哲学则是关于自然知识和社会知识的概括和总结[①]。"人的知识是在后天的实践中形成的。唯心论的先验论所谓先天就有的知识是根本不存在的。一个人的知识，不外由直接经验和间接经验两部分组成。间接经验，归根到底也是来源于直接经验。一切比较完全的知识，都是由两个阶段构成的，即感性知识和理性知识。感性知识是理性知识的基础，理性知识是感性知识的发展。它们是相互联系的，在实践的基础上统一起来，并随着实践的发展而不断发展。否定知识来源于感性经验，否定感性知识必须向理性知识发展，这就否定了认识论的辩证法。

科学知识对实践有伟大的指导作用，没有科学知识或

①《毛泽东选集》第3卷，人民出版社1968年版，第773—774页。

不以科学知识作为指导，就不能达到认识世界和改造世界的目的。“马克思主义就是共产主义从全部人类知识中产生出来的典范。”“只有用人类创造的全部知识财富丰富自己的头脑，才能成为共产主义者[①]。”要重视知识，坚持理论与实践相结合的原则，掌握知识，更有效地为革命和建设事业服务。

智力有智谋和力量的释法。《三国志·魏志·武帝纪》：“吾任天下之智力，以道御之，无所不可。”就是一例。通常，把智力叫“智慧”，人认识客观事物并运用知识解决实际问题的能力，集中表现在反映客观事物深刻、正确、完整的程度上和应用知识经验与从事实践活动中发展的。但又不等同于知识和天赋，它是先天素质，社会历史背景和教育的影响以及个人努力之方面因素相互作用的产物。智力是人的各种基本活动能力的综合，包括为观察力、注意力、记忆力、想象力、思维力等，其核心是人的抽象思维能力和创造性解决问题的能力。

智力的意义是心理学史上长期争论不休、至今仍众说纷纭的一个问题。最早给智力下定义的是德国心理学家斯特恩（L.W.Stem.，1871－1936），他认为智力是个体以思维活动来适应新情境的一种潜力。十九世纪来，法国心理学家比纳（A.Binet，　1857－1911）提出推理和解决问题的能力是衡量智慧的标志。二十世纪二十年代美国《教育心理学家》杂志开辟专栏探讨智力的含义和性质，许多心理学家发表了见解，最后仍未能得出统一的结论，但大致可以归纳为三种看法：推孟（L.M.Terman，1875－1956）等人认为智力

①《列宁选集》第4卷，第347，348页。

主要指抽象的思维能力；迪尔伯恩（W.F.Dearborn）等人则主张智力是学习的潜能；平特纳（R.Pintner）等人认为智力是适应新情境的能力。总之，我们可以把智力视为人的一种能力，就是人们在认识世界和改造世界实践中所表现出的身心力量。就是人们认识世界的能力和运用这种认识去能动地改造世界的能力。也就是人们的能力中与获得知识和运用知识相关联的那种能力。这就是人们常认为的一般能力。

技能是通过练习而形成的顺利完成某种任务所必须的活动方式或心智活动方式。技巧是在技能的基础上形成的熟练的、经过反复练习达到自动化了的活动方式。技能和技巧总是表现为对一定知识的认识和应用。如写字，凡对笔划、笔顺和笔法等语言知识的应用，从而形成书写技能和技巧。计算，是对数学、概念、法则、定理和公式等数学知识的应用。从而转化为具体的运算或论证技能和技巧。技能和技巧的形成又是以一定的能力为前提的。即一个人的能力水平和能力的个性差异直接影响个人形成技能和技巧的难易。所以，任何技能和技能的掌握和调节，都离不开与之相应的知识和能力。技能按其性质的特点，又分为动作技能（如打字技能、游泳技能）和心智技能（如写作技能、运算技能）。

人的知识、智力和技能，就存在形式来说是一种主观性的东西，是潜存于人脑之中的。但是，它是人们在实践中获得，并通过实践表现出来的认识，能力和活动方式。实质上，通过实践表现出来之后，就转化为客观性的东西，就成为一种客观存在。二十世纪初期，教育测量在美国盛极一时，堪称异军突起，所向披靡。它的理论根据是桑代克的著名信条，即所谓的“凡是存在的都有数量，凡有数量的东西都可测量。”

只要是客观存在的东西，人们总能够觉察它，感知它。任何不能表现、显露出来的知识、智力和技能，是根本不存在的，任何不能被感知的表现物也是不存在的。否则，便存在着人们无法认识的“彼岸世界”——不可知论的臆测。

从教育心理学、遗传学来看，人不同个性的知识、智力和技能是有差别的。这类差别就是我们测量要做的工作。人们观察出个体的知识、智力和能力的差别。对它的充分显露现象，用次序的等级或大小不等的数字进行量化时，这就是在对人的知识、智力和技能的测量。总之人的知识、智力和技能是可以测量的。

三、考试是一种特别的测量

考试是一种测量，是一种特殊的测量。它与其它测量相比，有自己的特点。从测量方式看，考试采用的是主测者提出问题，被测者回答问题的“回答”法。从测量对象方面看，考试的对象是人，它所测量的内容是人的知识、智力和技能。这就是考试作为一种测量的特点。成功的考试为被考者创设了一个对所有同时被考者一视同仁的情境，使不同被考者在同一条件下，用最明快的方式表露同一方面的知识和智能。这个特殊的情境，就是主考人提出问题被考者作答的考场。表露知识和智能的方式，就是主考人提出的考试方法。因此，各被考者所表露的内容，就具有量的可比性。由于被考者回答问题时，被考人或主考者对回答过程和结果作出同一规格的记录。从而最后的评等判分亦即考后的数据处提供了事实的依据。根据考试的目的、作用可以看出为了实现考试的目的，考试本质是一种测量。

当然，考试并不等全测量。也不是测量人的知识和智能的唯一方式。检验真理的唯一标准是实践，考试不是检验不是鉴定一个人的知识和智能的唯一理论和方法。但考试是人们迄今创造的测量人的知识和智能所用方法中，比较客观、比较公正、比较准确、效率较高的一种方法，它将被最为广泛地利用和研究。

第三节　考试的功能

考试是对人的知识和智力的测量。人们组织和实施这种测量，总是为了实现某种具体的目的。考试能够实现多方面的目的，是因为考试这种测量具有相应的多方面的功能。下面以“三论”观点看考试功能。

一、系统的观点

教育是一个系统，考试也是个系统。就教育这个系统来说，考试则是它的子系统，是一个不可缺少的要素。它与教育其它要素是有机联系不可分割的整体。可以讲，没有考试，教育这个系统就不会成为一个系统。因此，那种企图想取消考试办教育是不科学的。另一方面，考试本身还是一个系统，它的自身也是元素及其关系的总和。

考试系统（E）={考试的目的（A）、考试的作用（B）、考试的方法（C）、命题（D）、考试组成（E）……}

按考试系统和层次性讲，它可以分为：阶段考、综合考、全面考、断面考、定题考、泛考等形式。考试系统不仅有层次性，而且还有开放性和动态性。我们要不断地更换考试内

容，以适应不同时期对考试本身的要求。使它从不平衡——平衡——新的不平衡——新平衡。这样反复下去，以至无穷。保持整个考试系统的生机。比如，各种水平考试、资格考试、标准考试都在经常性地进行调整合格分数线和增补考试内容。

二、信息的观点

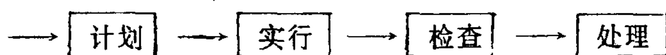
考试是一种信息反馈，学校教育工作是否有效，其关键之一，就在于是否有灵敏、正确、有效的信息反馈。衡量信息反馈的灵敏、正确、有效三者的程度，分别可以用时效、信度和效度来表示。这也是信息反馈的质量指标。所谓效度，主要是指这些感受来的信息经过分析后，有可能转化为指挥中心强有力的行动。信息反馈的效度首先决定于信度和时效，信度很低或没有信息反馈，是失真的或虚假的。指挥中心根据这种信息反馈进行决策，必然要造成失误。有了信度没有时效，则事过境迁。后来的信度也失去了存在的价值。损误时机，因而也不可能产生效度。信度不等于效度。信度反映信息、反馈的真实性，而效度则反映信息反馈的深广度，没有一定深广度的信息反馈，即使有信度，也不能为指挥中心提供高瞻远瞩、深谋远虑的决策。从学校管理来看，平时学生的作业、课内外提出的问题和反映等，都是各种灵敏、正确和有效的信息反馈，是不可忽视的。但是，这些信息反馈代替不了考试这一信息反馈。因为考试是教学过程中教和学这对主要矛盾和其它矛盾运动发展而显示出来的一种阶段性的表现，在这个阶段中，学生掌握知识和能力方面，不仅有量的积累，而且还有相应的质的转化，考试时、学生必须在规定的时间内，用一定速度（或熟练度）来独立思考和解答

问题。因此，一般来说，考试可能提供信度和效度更高的信息反馈。由于考试时间是固定的，为此，对考试来说，可以不考虑时效，而把信度和效度作为重要的品质指标。要是考试不能提供高效度和高信度的信息反馈，那么这样的考试也就失去了存在的价值了。

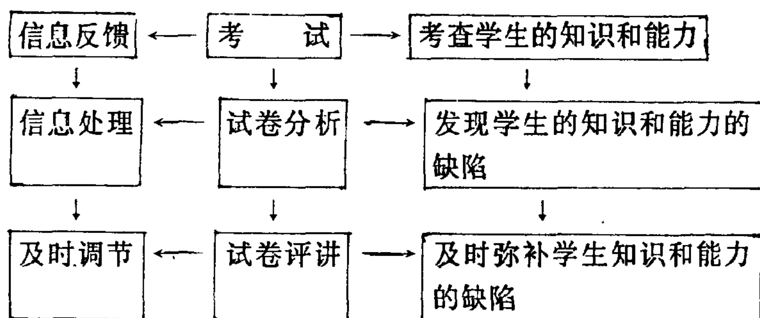
三、控制的观点

对于考试的控制，其指导思想应该是把考题作为获得应试者的信息的手段。应试者参加考试，其所答之考卷，就是主考者获得的反馈信息：鉴定质量。教师要进一步改进教学手段与方法，就需要通过考试获得及时、准确、可靠的反馈信息，以及对这些反馈信息的正确分析。依靠这种反馈，及时调整改进教学内容及手段，以达到不断提高教学质量的目的。过去对考试的控制的重点往往放在出题、施试、评分上。对考试结果的分析缺乏细致科学的指导研究。考试控制不仅重视成绩的评定，而且还应注重考试的结果分析。即分析应试者的反馈信息，从而为评价应试者提供可靠的依据。

考试的控制应采取如下程序：



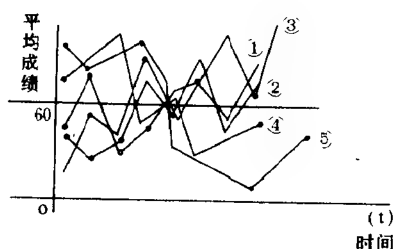
从控制论的角度，考试→试卷分析→试卷评讲，这三项工作是考试控制过程中不可分割的组成部分，它们之间的关系图解如下：



第一阶段：计划根据不同的考试目的，选择和确定考试题目，如考查学生对全部教学内容中基本概念的掌握情况，可选择若干概念性的小题，作为简答的问题，利用大面积随机抽样的办法，题目要简要，题目的范围要宽；考查学生对于某些原理掌握的情况；应出一些问答题；要求回答某些问题的基本原理，以考查学生对问题掌握的准确度和深度。再如，要考查学生利用所学的知识解决实际问题的能力，并在一定程度上考查学生智力和能力，就要出一些联系实际的问题。要引起学生的思考，等等。

第二阶段：实行考试中，为取得可靠的反馈，需要严格考场纪律，防止作弊及暗示等。

第三阶段：检查进行阅卷并进行试卷分析。从质量上进行深入细致的研究，从中找出问题，确定问题的性质，找出主要问题的主要原因，作为控制点，以便制定对策，进一步提出改进教学的措施。比如：通过试卷平均分数的统计，按时间顺序，给出纵观考试结果的质量动态图，以观察、分析教学质量的高低及稳定性。如图所示：



①随机波动型，成绩曲线在60分上下。随机波动，这表明该学生学习状况明显变化，维持在好、中、差的不同水平上。

②相对上升型，成绩曲线后阶段状态，略高于前阶段状态，前阶段曲线明显位于60分线上下，后阶段曲线相对上升并维持在60分线上下波动。这表明该学生学习有相对进步。

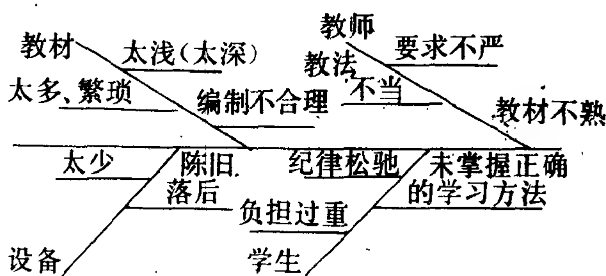
③显著上升型，成绩曲线后阶段状态显著高于前阶段状态。前阶段曲线显著低于60分线，后阶段却大幅度上升处于显著高于60分线的状态。表明该学生学习有显著进步。

④相对下降型，成绩曲线后阶段状态略低于前阶段状态。前阶段曲线显著高于60分线，后阶段曲线却在60分线上波动徘徊，表明该学生学习略有退步。

⑤显著下降型，成绩曲线后阶段状态显著低于前阶段状态，前阶段曲线明显高于60分线，后阶段曲线显著低于60分线，表明该学生学习有明显退步。

第四阶段：处理、肯定、巩固成绩。定出新的制度或改

进措施，找出尚未解决的问题，准备在下一个新的考试控制过程中进一步解决。比如：在三阶段中对试卷平均分数的分析后，应进一步分析形成上述情况的动态原因。一般讲，决定教学质量的因素有四个方面，即：教师、学生、教材、设备。在分析教学质量动态时，应找出影响教学质量的各种原因。然后按层次加以整理，绘制出原因分析图－鱼刺图。



从图中的线索找出影响教学质量的主要原因，在多种矛盾中 ，找出主要矛盾，以便确定管理重点，判定对策，提出解决问题的具体措施，使通过考试暴露出来的问题跃然于图上，能帮助主考者对问题的了解条理化。提高控制的准确性，为执行处理第一阶段的工作做好准备。

综上所述，考试这种测量，对于人的知识和能力具有评定功能，对于人的能力发展倾向具有预测功能，对于人的知识和能力增长过程具有诊断、反馈和激励功能。诸上功能不是分割的，而是综合起作用的。

第四节　考试目标、内容和考试标准的制定

考试是一项工程，在实施之前必须进行设计。它是属于计划工作。没有计划的考试是盲目的。考试设计是考试工作的首要环节。考试工程一般包括三大要素：考试的设计，考试的组织实施，考试成绩的使用。考试设计有三大基本要素：一是规定考试目标，确定考试内容和拟定考试标准，解决“考什么”的问题；二是决定考试方法和选择考试类型，解决“怎么考”的问题；三是编制命题计划，即将“考什么”和“怎么考”的规定变为实际工作的蓝图。

一、考试目标的规定

考试目标首先要根据考试目的，凭借考试的功能来规定，其次，就是要根据主考者和考试成绩使用者的关系来规定。大凡规定考试目标的依据有二：一是考试的功能。考试能够发挥什么功效，能够提供什么样的信息，实现什么的要求，这是规定考试目标的内在依据；二是客观需要。人们根据工作的需要提出对考试的期望，期望考试提供什么情况、信息。这是规定考试目标的外在依据。当然，考试目标的确定不是根据上述哪一种因素，而必须是二者的客观地结合起来。使得考试所能实现的，考试所提供的，恰是我们所需要的。定出的考试目标必须符合实际需要和实际可能实现才是考试的真正目标。否则，就不符合考试应该具备的要求。

规定考试目标，在行文上应尽可能明确、具体，切忌空

泛、太抽象，特别成绩的使用者是考试委托者，而考试的设计和组织实施者则是考试的被委托者。如高等教育自学考试，又如高考，规定：“选拔最有发展潜力的高中毕业生”为目标，就有空泛、抽象，不便于理解，不易于把握。有的学者认为：考试的目标应改为“选拔中学基础知识掌握较好，更能适应大学某类专业学习的高中毕业生。”，才便于确定考试内容，分类和选择考试方法，易于把握。

二、考试内容的确定

考试内容的确定，首先要研究目标的具体要求，研究考生的具体特点，哪些知识和智能是考生的共同优势，哪些是考生的薄弱环节，哪些在考生之间有较大差异。其次要研究选择的内容是否能够充分体现考试目标所要求的具体知识和考生的智能差异，足以反映考生的真实水平，使考试这种测量能有效地反映考生的成绩和区分增大。

考试的内容和考试的目标是有一定关系的。其关系反映在直接与间接两个方面。直接关系就考试目标本身就明确规定了考试的具体内容。例如，教学效果的单元考试和课堂测验等。间接关系就是考试目标并没有直接表示主要考的具体内容，而考试的具体内容则是根据考试目标指导下，再列出细目，加以具体化。例如高等教育自学考试，根据“从广大自学人员中选择具备高等教育学习水平的各类专业人才，以弥补普通高等学校和成人高等学校培养人才之不足。”的目的和当前自学考生的特点，必须分科或综合考核高等教育阶段的各门基础课。必须根据大学阶段的专业课的要求，加强考核某些课程，并突出某些能力的考核。

当然，有某些考试目标难以直接推出考试内容或者考生的具体特点不甚明了的考试，为了准确选择考试内容，常常需要先进行预试，或者认真分析、总结统计同类考试的研究成果，对不同方案进行比较研究，以求得对考试内容的最佳选择。

华东师大教育系试题分析小组，对上海市1983年高考、中专入学考试试题的统计分析表明：试题选定后，试卷分析应在考试之前，对预测试题所做的工作，经过分析可以筛选出一批比较好的题目，以提高正式考试试题的质量[①]。1984年，《人民教育》杂志开展了改革高考分类，科目设置与计分比例讨论，参加讨论的同志认为："考试科目、内容偏多，加重了考生的负担，考生穷于应付，不利于培养能力；发展智力；不利于考生发挥自己的特长；也不利于国家培养、选拔有特长的人才[②]"提出高中毕业实行会考制度，在省、市、自治区高中毕业会考的基础上，进行高等学校入学考试；高等学校入学考试分类加"细"，科目减少。

三、考试标准的制定

考试标准是编制试题和试卷、阅卷给分的基本依据，也是考生备考的基本依据，它是根据考试目标规定的关于考试范围、深度等方面的具体要求。就抽象的角度讲，考试内容和考试标准的关系，就象教科书与教学大纲的关系一样，教科书反映了教学的具体内容，教学大纲反映了这些内容的具

①载《华东师范大学学报》（教育科学版）1984年第4期。

②《改革高考分类科目设置与计分比例讨论》结束语，载《人民教育》1984年第7期。

体要求。

考试标准的特别重要任务就需要指明达到“合格”水平的基本要求。规定出“合格”的标准线。如自学考试、高考等。合格的基本要求是什么，在考试标准中须有具体说明。“合格”水平的标准一般具体体现在考试成绩——分数上。比如：托福考试合格成绩是550分。学校学年成绩考试一般是60分。但有的必须全部通过即为合格，不能用通过百分之几来表述，如驾驶员能力考试。

第五节　考试方法和类型的选择

一、常用的考试方法

考试方法是指考试的类型及要求考生回答问题的方式，按考生回答问题的方式区分。考试有口试法、笔试法、操作法、特殊问题处理法。其中，笔试法又分闭卷笔试和开卷笔试。

口试法：就是在主考者的面前，考生用口述的方式回答问题，主考者根据回答的正确程度加以评分。

口试法的优缺点是：其优点是主考者能够通过连续多问及时搞清回答中表述不清的问题，试题取样宽，较灵活，从而提高考查的深度和清晰度。能够考查出考生对知识掌握的牢固、熟练程度，能够观察到考生的外表、风度、言谈的格调和在外界压力下的承受能力。口头表达能力以及思维的敏捷性。降低了作弊的可能性，提高考试的真实性。其缺点在于不能对考生进行群体同时考查，耗费时间太多，效率较低。

增大考生的思想压力，考试的成绩常受与考试目标无关的因素的影响，如口齿伶俐与否，精神的兴奋程度等。而往往有由于口齿清楚、反映灵敏、精神振奋，在回答问题时给评判者以良好印象，予掩盖其知识和智能上的不足，评分不够客观、准确。

笔试法：就是让考生在事先编印好的试卷上笔答。主考者根据考生试卷解答的具体情况进行正确的评分。

笔试法的优缺点：其优点在于考试的取样大，一次考试能够出十几道乃至上百道题的问题，对知识和智能的考查其信度和效度远较口试为高；考生容量远大于口试法。费时少，效率高。同时对考生的心理压力相对于口试法减少，较能正常发挥考生的水平。它能保留考生回答情况的真实材料，评分也较口试客观。其缺点在于考生舞弊机会增多，考生在回答问题时有较多的凭借猜测、欺骗、作弊而取得分数的机会。考查思维的敏捷性、口头表达能力不如口试法。

操作法：是主考者让考生操作由主考者设计好了的考试题目。以实际操作技能中的表现对考生的知识和智能加以评分。

特殊问题处理法：实际上是笔试、口试和操作法的综合运用。

上述的各种考试方法，各有其特点，各有其长处，分别运用于不同目标、不同的内容的考核。笔试不能代替口试和操作考试。首先它们在能力的检查上有不同的侧重点。 笔试只考动脑的能力，而在口试和操作考试中，考生不仅要动脑，而且要动手。目前我国中小学考试主要弊端之一就是几乎一律笔试，看不出动口动手能力如何。因此中小学生普遍

存在着一种不良倾向：不重视培养学生动口和动手的能力，仅仅满足于笔答几个试题。其次，在智力品质的要求上有不同的特点。笔试主要考学生思维的准确性，深刻性和广阔性。而口试和操作考试除了要求学生思维具有上述三种品质外，还分别侧重于考生的思维敏捷性和精细性。再次，在心理状态上有不同的特点。笔试答题时间一般长，题目多，考生有一定的回旋余地，因此心理处于紧张状态的时间较短暂，而口试一般时间短，题目少，考生回旋余地少，而且要充分估计主考者可能出的追加问题，同时又要考虑口头表达等因素，考生心理紧张程度较高，而且持续的时间也较长。笔试法是现代考试中应用最为广泛的一种方法，研究此种考试方法的问题较多，以下各章主要论述它。

二、常用的考试类型

考试类型的划分方法有许多，最基本的方法就是根据考试的目的、作用和功能来进行划分。

按考试的作用，可将考试划分为成绩考试、水平考试、学能倾向考试和诊断考试。

按给分的客观程度划分，有主观性考试和客观性考试。

按考试的目标、试题和成绩是否经过标准化处理，可将考试区分为标准化考试和非标准化考试。标准化考试，按具体标准和反映分数的方法又可以分为常模参考考试和目标参考考试。

按考试试题类型划分：有论文法考试，判断选择法考试，计算法考试。

按考试规模划分有课堂考试和社会性考试以及国际性考

试。

按考试的内容范围划分有单科考试和综合考试。

按考试的社会性划分有学校成绩考试，自学考试等等。

三、考试方法和类型的选择

考试方法和类型的选择首先根据考试目的确定是属于成绩考试还是水平考试或是学能倾向考试，或是诊断考试等，如高考，根据"选择中学基础知识掌握较好，更能适应大学专业学习"的要求和当前高中毕业的特点，必须分科或综合考核高中阶段的各门基础课；必须根据大学阶段的专业学习要求加强考核某些课程，并突出某些能力的考核，就可以确定为学能倾向考试和成绩考试的综合。又如高等教育自学考试，目标是考核自学应试者的学历水平，对达到专科或本科毕业水平的，发给相应的学历证书，因此可以确定是水平考试。

其次选择考试方法。如高师自学考试，就必须利用口试考语言表达能力，实验操作和笔试。如学校成绩考试，除语言专业的语言、听力、对话是口试，实验、课程设计、论文答辩是典型问题处理外，其余都是笔试。

再次根据考试规模、内容和标准，选择试题的类型，如少量的高级考试，考生数量少，每科只考一次，要求考出考生的潜在能力和创造力，最好选用主观考法。如高考，学生数量大，无平常测验和作业检查，要求题目量可大，而又分客观、快速，因此，除个别学科外，应普遍选用或主要选用判断选择题，采用标准化命题考试。

考试目标、内容、标准考试方法，与类型有密切的关系。

就考试类别的划分，有的是根据目标、内容等特点进行的，当考试目标、内容等确定后，只需要加以识别，认定，就能确定考试类型。但有的则不然，它需要对于题型的斟酌，有的考试题型的选择，不同于对于根据其他的特点划分的考法的选择，具有较大的灵活性和随意性。因此，需要专门的研究，探讨出所需的类型的考试方法。

第三章　考试科学性的评价指标

考试的科学性包括考试形式与考试内容的科学性两个方面，科学的考试是二者的辩证统一。判断考试科学与否，必须用评价指标进行度量。考试的评价指标是考试质量分析的依据，是编制试卷基本依据。就考试发展的现状看，评价考试科学性主要用信度、效度、区分度和难度来评价。

第一节　考试的信度

一、信度的概念

对于考试来说，信度反映了考试结果的的可靠程度，反映考试对象在考试前后表现的一致程度。这就好似一项有效的测量，如果能对同一对象实验多次，多次所测的结果应该是比较一致，比较稳定的。一致性程度越高，稳定性越大，这项测量越可靠、越可信。信度的常识性定义是用测量工具测量同一事物前后一致的程度。它是反映测量结果的稳定性、可靠性的一个指标。一项考试如果针对同一组考生实验多次，所测得的各组成绩之间的一致性程度越高，考生组在多次考试中的成绩排列越稳定，这项考试也就，越可信。显然，对同一组考生，考试成绩的稳定性程度，多次考试所得成绩的一致性程度，是衡量考试质量高低的重要指标，这个指标，

我们称为考试的信度。

如果一次考试的结果十分准确，完全没有误差，成绩与考生的真实水平完全一致，那么这个考试的信度就最大，即这次考试的分数完全可靠。信度的高低主要反映在考试过程中随机因素影响的大小，即存在随机误差的大小。

通常，把考试的信度定义为：一项考试的信度，就是这项考试的一组成绩和对同一组考生实施多次考试所得的另一组成绩的一致性程度。

信度也是个统计学上的概念，可以用相关系数来表示，我们称之为信度系数。相关的程度越高，信度也就越高。换言之，信度就是观察值（即考试考生实际所得分数）与真值（考生完成这次测验应该得到的分数）的相关系数。用数学式子表示信度则为：在实际分数的方差中，随机误差所占的比重。

即：
$$r_R = \frac{S_x^2 - S_e^2}{S_x^2}$$

在上式中，S_x^2是总变差或实得分数变差；S_e^2是误差变差。

考试信度可以表示为考试分数的方差与考试误差的方差之差，与考试分数方差之比。

r_R的最高值为1，表示考试完全反映了考生的稳定水平；最低值为0，表示考生得分完全是随机性的，而与考生本身的水平无关。

式中，$S_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n}}$ 是考生分数的标准差。S_e是分数误差的标准差。

X_i是第i个考生的考试分数。 $\overline{X}=\frac{\sum_{i=1}^{n} x_i}{n}$表示考生的平均分数。n为考生人数。

我们将代表真实分数个体差异情况的差异量数和代表考试分数个体差异情况的差异量数进行比较。

令$x_i=y_i+e_i$（i=1、2、……，n），e_i是第i个考生分数的误差。y_i是在等价考试中第i个考生的分数，则

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}x_i=\frac{1}{n}\sum_{i=1}^{n}y_i+\frac{1}{n}\sum_{i=1}^{n}e_i=\overline{y}+\overline{e}$$

$$S_X^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{X})^2=\frac{1}{n}\sum_{i=1}^{n}(y_i+e_i-\overline{y}-\overline{e})^2$$

$$=\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2+\frac{1}{n}\sum_{i=1}^{n}(e_i-\overline{e})^2$$

$$+\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})(e_i-\overline{e})$$

由于误差，e_i是x_i偏离真值y_i带来的，与y的分布无关，因此上式最后一项为零。于是

$$S_x^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2+\frac{1}{n}\sum_{i=1}^{n}(e_i-\overline{e})^2=S_y^2+S_e^2$$

令$r'_R=\frac{S_y^2}{S_x^2}$，它能反映考试成绩分布和真实成绩分布的接近程度，亦即考试的准确度。

考试信度反映的是考试客观性，考试结果准确性的程度。但是，考生的真实分数及其方差，我们是不知道的，如果知道了，也就不需要考试了，至少说可以不用考试这种方法了。然而，我们可以对一组考生，在相同条件下，进行大量的同类考试，所得的分数的平均值就十分接近真实分数。为此，

信度在某种程度上还可以定义为：一项考试的信度，就是这项考试的一组成绩，与对同一组考生实施等价考试所得的另一组成绩的相关系数。

根据相关系数的计算公式，信度系数可用下式表示：

$$r=\frac{\sum_{i=1}^{n}(x_i-\overline{X})(y_i-\overline{y})}{nS_xS_y}$$

由上式可见，r只是一个系数，它的值总是在1.00－－1.00之间。r＝1.00表示两次考试的两组成绩完全等价，完全一致，r＝－1.00表示两组成绩的排列恰好相反，称完全负相关。

二、信度的种类与计算方法

评卷员信度，一份考试试卷施考后，将学生的答案让不同评阅员阅卷给分，若不同评阅员所打的分数大致相同，这个测试卷便有高信度。反之，信度则低。客观性的试题评分准则客观、准确，不论那个评卷员或那部机器去评分，所得结果一致。所以评卷员的信度很高，非客观性试题评分准则有弹性。评卷员渗杂着自己的情感判断打分，不同评阅员便会由于不同观点和“准则”而评得不同结果。所以难以保证有较高评卷员信度。改善非客观性试题的评分准则，就是要使它能达到最大的客观信度。非客观性测验的评卷员信度是可以提高的。

克朗巴赫（Cronbach）在1951年发表的《α系数和测验的内在结构》一文中提出作为估计内在一致的可靠性指标的α系数分式：

$$\alpha=\frac{k}{k-1}\left[1-\frac{1}{S^2}\sum_{j=1}^{k}S_j^2\right]\quad(0\leqslant\alpha\leqslant1)$$

评卷员信度可用Cronbach系数法来建立评卷员信度的度量计算。式中k是试卷所含的题数，S_j^2是第j题得分的方差。S^2是考试总分数的方差。即有

$$S_j^2=\frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\overline{x_j})^2\qquad S^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y_j})^2$$

n：考生人数；

X_{ij}：是第i个考生第j题的得分数；

$\overline{x_j}$：是考生j题的平均分数；

y_i：是第i个考生的总分数；

$\overline{y_i}$：是考生总分的平均分数。

当项目数k较大时，可以保证$0\leqslant\alpha\leqslant1$。式中 $\sum_{j=1}^{k}S_j^2$ 反映了每一个阅卷小组成员在各题的阅卷中离异程度，而S^2反映了各阅卷小组成员在评阅整体上的离异程度，它当然受各题的评阅的离异程度的影响。因此，$\sum_{j=1}^{k}S_j^2/S^2$这一项既反映了在对评卷系统的每一题的评阅过程中的随机因素的大小，又反映了各个评卷员的整体的评阅结果中的随机因素的大小，当问卷的标准答案越明确，量化标准越准确，各个评卷员对评卷的评分方法和评分标准的规律越一致，评阅越客观，随机因素越小。$\sum_{j=1}^{k}S_j^2/S^2$项越小，而α则越大，反映评阅工作的可靠性和稳定性就越高，因此，对α系数公式进行以上方法处理，可用作估算评卷员的信度，α系数即信度系数。

用Cronbach系数法计算时，有多少个评卷员，就要有多

少列数据，因而要用多列相关公式。多列相关公式计算比较复杂，在实际计算中常把分数转换成等级形式。由下列公式表示：

$$\omega=\frac{\sum_{i=1}^{n}R_i^2-\left(\sum_{i=1}^{n}R_i\right)^2/n}{k^2(N^3-N)/n}$$

其中W为评卷员信度系数，k为评卷员人数，N为被评试卷数，R_i为第i份试卷得等级之和。

计算评卷员信度的例子：

学生编号	评卷员					
	A	B	C	D	E	总分
1	2	1	2	3	3	11
2	2	3	3	3	3	15
3	2	1	2	2	2	9
4	3	3	1	4	2	13
方差	0.25	1	0.5	0.5	0.25	5

评卷方差和 $\sum_{i=1}^{n}R_i^2=0.25+1+0.5+0.5+0.25=2.5$

考生总方差 $S^2=5$

评卷员信度 $\alpha=\frac{1}{5-1}\left[1-\frac{2.5}{5}\right]=0.625$

第二类，再测信度r，所谓再测信度即用同一个测试工具在两个不同的场合施行相同的考生而求其结果的相关，就能获得相关系数。这一相关系数称为信度系数。即是说，用考试对考生施考后，过一段时间再考一次，将两次考试的结

果加以比较，求得两次分数之间的一致性。再测信度的大小用两组分数之间的相关系数（相关系数表示具有一定关系的两组数据的关联程度或一致性的一种指标）表示。

计算再测信度系数的公式如下：

$$r=\frac{\sum_{i=1}^{n}(x_i-\overline{X})(y_i-\overline{y})}{nS_xS_y}$$

$$=\frac{n(\sum_{i=1}^{n}x_iy_i)-(\sum_{i=1}^{n}x_i)(\sum y_i)}{\sqrt{[N(\sum_{i=1}^{n}x_i^2)-(\sum_{i=1}^{n}x_i)^2][N(\sum_{i=1}^{n}y_i^2)-(\sum_{i=1}^{n}y_i)^2]}}$$

其中，$\overline{x}$、$\overline{y}$分别为第一、二次考试分数的平均数；S_x、S_y分别为第一、二次考试分数的标准差；r的最大值为1，表示最高再测信度。而最小值为0，表示最低的再测信度。若再测信度高，则考试成绩稳定，不受两次考试期间所发生的变化的影响。反之，则考试成绩不稳定，受两次考试期间所发生的变化影响。

计算再测信度的例子：

学生编号	第一次考试分数x_1	第二次考试分数x_2	x_1^2	x_2^2	x_1x_2
1	2	3	4	9	6
2	4	3	16	9	12
3	5	4	25	16	20
4	6	6	36	36	36
5	7	7	49	49	49
总分	24	23	130	119	123

信度系数 $r=\frac{5\times123-24\times23}{\sqrt{(5\times130-24\times24)(5\times119-23^2)}}$

$$=\frac{63}{\sqrt{74\times66}}=0.90$$

再测信度有一个不足之处就是不能同时对考生进行考试。时间的间隔可能会影响信度。因此，我们提出用平行试卷信度来弥补这一缺陷。把一个考试编制成两份性质相同而信度不一样的试卷。然后分别考试学生。计算成绩后，以相关统计求得相关系数，即为信度系数。这个平行试卷信度值的大小所展示信度的高低及其计算方法与上述的再测信度系数无差别。若以用一试卷给同一批考生再试两次，考生可能对第二次相同试题的考试产生心理抗拒，以致考试成绩不能真实反映考生的水平。故此以两份文字不同的平行试卷考试学生，可减低学生的上述心理抗拒，因而提高考试信度。但是，若平行试卷设计不完善，不能保证两份试卷的性质相同，考试题的文字性质的改变会影响考生的真实表现，造成考生两次考试成绩不一致，反而降低考试信度。这是试题设计者需要特别留意的。

第三类，分半信度，为克服上述由两次考试所造成的困难，可在一次施考的情况下，通过某种手段将一份考试题目分成相等的两半（常用的方法是按奇数与偶数题分半），并将这两部分的分数分别统计，计算其相关作为信度指标，称分半信度系数（或内部一致性系数）。

由于分半法是将一套试题分作相等的两半，把一次考试看作题目减半的两次考试，而考试的信度是与试题数量密切相关的，因此分半信度比整个考试的实际信度要小一些。用

分半法求得的两半分数的相关，只是半个考试的信度，因此必须用斯皮曼——布朗分式（Spearman－Brown），一项考试的信度 γ 与它的分半信度 $\gamma_{半}$ 有下列关系：

$$\gamma=\frac{2\gamma_{半}}{\gamma_{半}+1}$$

计算分半信度的事例：

学生编号	半份试卷分数 单数试题 x_0	双数试题 x_F	X_2^2	x_F^2	x_0x_F
1	1	1	1	1	1
2	1	1	1	1	1
3	1	1	1	1	1
4	1	2	1	4	2
5	2	2	4	4	4
总分	6	7	8	11	9

半份试卷的相关系数 $\gamma=\dfrac{5\times9-6\times7}{\sqrt{5(8-6\times6)(5\times11-7\times7)}}$

$$\doteq0.61$$

全份试卷信度系数 $\gamma=\dfrac{2\times0.61}{1+0.61}=0.76$

另一个重要的估计信度的公式是分半信度公式的推广，当一个考试是由两个等值的考试复合而成时，如果已知其中一个考试的信度为 γ_{xx}，整个考试的信度为：

$$\gamma_{kk}=\frac{K\gamma_{xx}}{1+(K-1)\gamma_{xx}}$$

用数学方法可以证明，这个公式反映了试卷长度与信度之间存在着递增关系，γ_{kk}是K的增函数，即增加考试的长度可以提高测验的信度。但不是说考试越长越好，一般情况下，考试的长度大约控制在一百个题左右。大量的统计结果表明，若进一步增加测验的长度，信度提高的幅度微乎其微，得不偿失。

第四类：试题的信度。将一份试卷施考一次。然后把每一个考生的各个试题所得的分数代入有关的公式，求得的数值，即为试题信度系数。这个系数的最大值为1，表示最高的信度，亦即全卷各试题的一致性最好。最小值为0,表示最低的信度,亦即全卷各试题的一致性最弱。计算试题信度有两种方法：一种是给考生在每一试题的得分只有1或0而计算。1分表示考　生在这一试题的答案是正确的。0分表示考生在该题的答案是错误的。另一种是为学生在每一试题的得分多于1或0两个分数，固而考生答案的正确程度获得不同的分数。前一种方法较适用于客观性试题，尤其是长问题解答。这一种方法应用顾拜尔——李才舜（KR_{20}）公式计算：

$$K_{R_{20}}=\frac{n}{n-1}\left(1-\frac{\sum_{i=1}^{n}p_iq_i}{S_x^2}\right)$$

其中，P_i表示第i个题目考生的通过率（通过率：即通过第i个题目的考生数与全体考生数之比）$q_i=1-P_i$，n表示考试所含有的题目数。S_x表示考生在该考试中所得分数分布的均方差。

后一种方法应用艾尔克朗巴赫（Cronbach）α系数计算，

$$\alpha = \frac{K}{K-1}\left[1 - \frac{1}{S^2}\sum_{i=1}^{n} S_i^2\right]$$

下面分别以上述两种方法展示计算试题信度系数的例子。

计算各项选择题试卷的试题信度系数的例子：

考生编号	试题编号 1	2	3	4	总分
1	1	0	0	0	1
2	1	1	0	0	2
3	0	0	1	1	2
4	0	0	0	0	0
5	1	1	0	1	3
6	1	1	1	1	4
总分	4	3	2	3	12
p_i	0.62	0.50	0.33	0.50	
q_i	0.33	0.50	0.62	0.50	
p_iq_i	0.222	0.250	0.222	0.250	

$$\sum_{i=1}^{n} p_iq_i = 0.222 + 0.250 + 0.222 + 0.250 = 0.944$$

考生总平均　即$\frac{\text{全体总分}}{\text{考生人数}}$，　$n = \frac{12}{6} = 2$

考生总方差　即$\frac{\text{考生个人总分与考生总平均之差的平方和}}{\text{考生人数}}$

$$S_x^2 = \frac{(1-2)^2 + (2-2)^2 + (2-2)^2 + (0-2)^2}{6}$$

$$+(5-2)^2+(4-2)^2 = \frac{10}{6} = 1.667$$

试题信度系数 $K_{R_{20}} = \left(\frac{4}{4-1}\right)\left(1-\frac{0.944}{1.667}\right) = 0.58$

计算长问题的解答试卷的试题信度系数的例子：

学生编号	试题编号 1	2	3	4	总分
1	4	3	4	4	15
2	2	5	5	5	17
3	3	5	5	5	18
4	1	3	1	1	6
5	5	5	5	4	19
6	4	3	4	4	15
方差	1.81	1.00	2.00	1.58	16.89

试题方差和 $\sum_{i=1}^{n} S_i^2 = 1.81+1.00+2.00+1.58 = 6.39$

考生总方差 $S^2 = 16.89$

试题信度系数 $\alpha = \left(\frac{4}{4-1}\right)\left(1-\frac{6.39}{16.89}\right) = 0.82$

信度系数是难以解释的，这些系数反映着计算它们的方法。各团体的变化，各次考试之间的时间，受考试的团体的特点以及其它条件因素等等的影响。但是，凯利CT.Ckelley，1927）却只为一个年级的信度系数提出下列最低要求，可供参考与借鉴。

为了决定一个班在一个学科或一班在学科上的地位，信

度系数要达到0.50；

为了鉴定一个班在两种以上学习线上的成绩，信度系数要达到0.90；

为了鉴定一个班在同一学科或一班学科上的地位，信度系数要达到0.94；

为了两种以上学习线上鉴定各个个体，信度系数要达到0.96。

三、提高信度的方法

提高考试信度，就是减少考试时产生的误差，增加考试的客观性和准确性。提高考试信度，是提高考试质量的重要途径和基本技术。

提高考试信度的基本方法有：

第一，提高试题的区分能力，准确反映不同水平考生分数的差异。

考试信度描述的是多次等价考试分数分布的一致程度。如果同一个考生两次考试的分数差异，相对地小于不同考生分数之间的差异，各考生的成绩大小排列的次序就不会打乱，就能够获得较高的信度。相反地，如果同一个考生的分数变动，相对地大于不同考生间分数的差异，考生成绩的分布状况就要发生大的变动，就会使考试信度下降。

客观性是信度的另一方面。当考试的分数没有受评分者的偏向或个人判断的影响时，这个考试就可以说是客观的，考试需要消除评分者的主观偏向或成见，这早就被公认为考试评分的一个因素了。很明显，如果一个考试由同一读者在若干种情况下评分，第一次却得到不同的分数，或者这个考

试由若干读者评分，每一读者都给予不同的分数，那么这个考试就很难说是高度可靠的了。如果选入一个考试之内的都是非常客观的项目，那么关于什么是正确的答案就会很少或没有异议了。

试题的区分能力，也叫区分度，它就是试题在用于考试时使水平高的考生得分高，水平低的考生得低分的倾向力。这一问题本章第三节中将讨论。

第二，适当增加试题的数量，扩大试题覆盖面。

当考虑一个考试的信度时，还有一个和信度有关的方面，就是“适度”，一个考试是有关一个人的知识和技能的一个样本。它不是用以评定这个人所有的知识和技能，它只是努力去获得足够的样本以便作出结论。这就告诉我们，为了取得高信度就必须注意取样的适度问题。编制一个考试，如果题目太少，便可能仅仅选取某一个考生所知很少的那些题目，而是漏掉他所知很多的那些题目，或取舍情形恰恰与此相反。因此，假使选取另一种样本（或给予另一种考试），结果也许根本不同，以致无法确定他对本门学科的全面知识。我们采用适当的增加试题的数量，使此试题的题目应大到可为全体的足够的代表。考试的准确性就要看取样的范围而言。因此，考试的信度就是长度的部分函数。

为了更进一步阐明上述观点，从斯尔曼—布朗的公式：

$$\gamma_n=\frac{n\gamma}{(n-1)\gamma+1}$$

式中，γ为试题数量较少的考试信度值，γ_n是题目增加到n倍后的考试信度值。如果一项由10道题组成的考试信

度值为0.5，将它的项目增加到95道，增加试题质量与原题相当，那么，题目增加后的考试的信度值为：

$$\gamma_{9.5}=\frac{9.5\times0.5}{(9.5-1)\times0.5+1}=0.905$$

这说明，成倍增加题量使考试信度明显提高。

若将增加到95道后再增至150道，增加后的信度值为：

$$\gamma_{150}=\frac{1.5\times0.5}{(1.5-1)\times0.5+1}=0.93$$

这说明，增加题量到一定数目的时候，再增加题量对信度的提高就很小了。

一般讲，我们有：

$$\gamma_n=\frac{n\gamma}{(n-1)\gamma+1}=1-\frac{1-\gamma}{(n-1)\gamma+1}$$

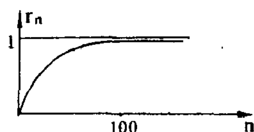
$\because 0<\gamma<1 \quad \therefore 1-\gamma>0$

若题目增加，则$n>1$。因此$\frac{1-\gamma}{(n-1)\gamma+1}>0$

$$\therefore \quad \gamma_n=\frac{1-\gamma}{n\gamma+1-\gamma}=1-\frac{a}{n\gamma+a} \qquad (a=1-\gamma)$$

这时我们就可以把γ_n看成是n的函数，n为自变量，γ_n为因变量，其图象为：

由图象可知：当n在大于某一数值（如此处为100）的范围内变化时，γ_n的变化率非常小。这就说明增加题目的数量一定要适当。



增加试题能够提高考试的信度，从内在结构看，实际上增加了内容，使得考题内容有更好的代表性，由此能减少由于抽样特殊而带来的测试误差。当然，增加题量、分布要合理，要尽量扩大对所考内容的覆盖面；否则，对于提高信度的作用不大的。增加题量，注意不要出偏题；怪题和没有考核意义的题目；否则，可能减小考试的信度。总之，增加题量，不降低题的质量，是提高考试信度的主要方法。

第三，尽量消除测试中的干扰因素，减小随机误差。

考试过程中，考生的过分紧张、疲劳，考生碰巧对某些有研究或根本无研究，题目用语不准使考生不知所云，特别是考场纪律松弛，考生有作弊行为，是对客观式考试的一种干扰，是考试随机误差产生的一个方面。努力消除这些因素干扰，是提高信度的重要措施。

第二节　考试的效度

一、考试效度的概念

考试信度是表征考试质量的重要指标，但它并不是衡量

考试质量的唯一指标。一项考试信度高，并不能完全保证它一定具有质量。反映一项考试实现其既定目标的成功程度的指标，是考试的效度。

效度是科学考试的必备条件，是考试所反映出与考试内容相对应的能力水平，并达到一定目的的程度，就是说考试要能够准确地测量出它要测量的成绩。考试之效度总是针对一定的考试目的或被测量的特性而言的。没有效度或低效度的考试是没有价值的。若一项考试的效度达到满意的程度，我们说这项考试是有效的。反过来，这项考试便是无效的。例如，算术考题对考试算术的知识与技能是很有效的。但对考试其它学科的知识来说，却是效能较低的。微积分考题用以考试大学理工科学生的知识与技能是有效的。对一般初中学生来讲则是无效的，一项考试对某一个目的是非常有效而对另一个目的却是完全无效的。这些事例使我们较清楚地理解效度这个概念。表面看来这似乎是个显而易见的标准，其实违反这个标准的错误情况还是不少的。常见的情况是编写一项考试某门学科的考题，因言语艰深难懂，只有那些能读懂的学生才能回答。所以，一项考试工具如果要有高效度就必须选用考生已经学过的材料。超过考试大纲要求的考题显然是违反效度标准的，用它来进行考试是无效的。总起来，效度是反映考试准确性，有效性的指标，效度的高低是要受条件误差的影响。反过来，效度的高低也说明考试过程中的随机误差所占的比重。随机误差所占的比重越小，信度越高。效度是由考试过程中的系统误差所占的比重决定的。对于一个考试系统来说，显然要有一个已被证明具有定效度的效标才能确定本次考试的系统误差有多大，比如用钟来测量时间，

如果我们已有一个标准钟，这个标准钟用于测定间是有很高效度的，那么对其它钟用于测定时间的有效性，就可以与这个标准钟相比较。没有标准钟，是无法估量一个钟表所测定的时间的有效性到底有多大。效度和信度一样，是一个相对概念，因此，不能只说某份试卷是否有效度。而应该讲此试卷在考那些方面是有效的。

效度原是个统计学的概念，它可以凭借一定的措施加以鉴定，鉴定效度有赖于寻找一种外在的，同这个考试题相关的标准。一个考试工具与某一外在标准的相关系数叫效度系数，相关的程度越高，该考试的效度就大。这自然是根据这样的基本思想，即那个标准化的工具可被认为是进行考试的准则，一般说来，那些考试大体具有经过实践证明的高效度，因而可作为标准。

二、效度的种类与计算方法

效度的种类。

效度有许多种，它们侧重的问题互不相同，最常遇到和使用的是内容效度、效标关联效度和结构效度。

第一类，内容效度，指施考内容与预定要考的内容间的一致性程度。即试题内容有无过难、过易或偏差，是否考了该考的内容或文字表达有无不当，不准确等等。由于这种效度主要与考试内容有关，故称内容效度。

内容效度的高低，除了要考虑考试蓝图的一致性和考试蓝图与试题分布的一致性外，试题的质量亦需要考虑。试题本身的缺点和正确答案的谬误都能影响考试效度的高低。因此，检定考试效度时亦要从试题质量方面分析。

第二类，效标相关联效度：

这类效度是研究考试结果同某种外在参照标准相关高低的一种经验的方法。它又可分为同期效度和预测效度，如考试实施目的在于评定现时的成绩，就要求同期效度，如目的在于考试学习某种专业或从事某种职业的未来成就，就要求预测效度。

同期效度，一个考试的同期效度是表示这个考试考核学生所得的分数与同时期的另一考试或准则评核同一班学生所得的分数的一致性。这种一致性是以两组分数的相关系数来表示，这个相关系数亦称为同期效度指数。若指数值为1，同期效度最大，若指数值为0，同期效度最低。例如举行高中毕业统考，目的在于检查学生的学业成绩是否达到了高中各科教学大纲规定的要求，也就是说是否达到了毕业的标准。要检验毕业统考的效度，可以用毕业班学生三年学习总成绩做效标而求其相关系数，这种系数所表示的就是同期效度。

预测效度，一个考试的预测效度是表示这个考试的成绩能够预测受试学生将来的某种特定行为或表现的程度。若预测的程度高，这个考试的预测效度便高。反之，它的预测效度则低。预测程度可以相关系数表示。这个相关系数可称为效度指数。具体的做法是用考试测量一批学生，得出一组考试分数。过了一段时间，把这批学生通过这个考试预测的行为或表现以某种量表或准则作评校。得出另一组分数，然后以相关统计求出这两组分数的相关系数。这就是这个考试的预测效度指数。若指数值为1，预测效度最大；若指数值为0，预测效度最低。

第三类，结构效度，结构效度指考试分数能够说明理论

研究所得的某种特质或结构的程度。如果一份考试题有结构效度则考试必然同预定的的结构变化相一致。比如说：我们要研究创造力的结构，应从理论上研究有创造力的人在作业方面和没有创造力的人有不同的结构。从而提出一种理论详细说明这两种人在行为上的差别，这样就可以通过观察个人的行为，断定哪些人具有创造力。

效度计算方法

内容效度的计算公式是郭朗伯（Cronbach 1971）设计的。他的做法是选设A、B两组对所考试的内容有认识的专家。每组分别编制两份性质相同，文字不一样的试卷。两组合共编制出A_1、A_2，B_1及B_2四份试卷。把这四份试卷给予同一批学生考试，评分后把两份卷的成绩作一相关统计。求得六个相关系数；A_1与B_1之系数$\gamma_{A_1B_1}$，A_1与B_2之系数$\gamma_{A_1B_2}$，A_2与B_1之系数$\gamma_{A_2B_1}$，A_2与B_2之系数γ A_1与A_2之系数$\gamma_{A_1A_2}$，B_1与B_2之系数$\gamma_{B_1B_2}$。最后，把这六个系数代入下列公式：便求得效度指数：

$$效度指数\frac{\gamma_{A_1B_1}+\gamma_{A_1B_2}+\gamma_{A_2B_1}+\gamma_{A_2B_2}}{2(\gamma_{A_1A_2}+\gamma_{B_1B_2})}$$

若指数值为1，则效度最高。若指数值为0，则效度最低。

预测效度指一个考试对处于特定情境中的个体的某些我们感兴趣的行为进行预测的有效程度。它以那些被预测的行为表现作效标。一个考试预测得越准其有效程度就越高。其计算方法为：

①求效度系数：这是最常用来建立预测效度的方法。即求考试分数与效标分数的相关系数。

$$\gamma_{XY}=\frac{\Sigma xy/n-(\bar{x})(\bar{y})}{S_xS_y}$$

其中，γ_{xy}为效度系数，x、S_x分别为考试分数的变量和标准差；y、S_y分别为效标分数的变量和标准差。

这种方法便于比较，但只提供了相关系数大小。究竟多高为可接受，还未定论。一般为相关显著即可。此外，在用此法前，需要先判断考试分数与效标分数之间是否为直线关系，若否，必须采用特殊的计算方法。

②分细法，选按效标分数分组，然后看每个组问题预测源的分数差异是否显著。

$$t=\frac{\bar{x}_s-\bar{x}_n}{\sqrt{S_s^2/n-S_n^2/n_n}}$$

其中，$\bar{x}_s$，S_s，n_s分别为成功组（学习优秀）的预测源的平均分，预测源分数的标准差及人数；$\bar{x}_n$、S_n，n_n分别为不成功组（不胜任工作和学习）的预测源的平均分，预测源分数的标准差及人数。

然后查t检验表，若差异显著，该次考试有效。

③取舍正确性：当考试用来作取舍的依据时，其有效性的指标就是正确决定的比例。有两个指标可用：

$$P_{cT}=\frac{\text{命中}}{\text{命中}+\text{失败}}=\frac{\text{命中}}{\text{总决定数目（N）}}$$

命中包括成功地接受和拒绝；失败包括错误地接受和拒绝。

$$P_{cq}=\frac{\text{成功人数（成功地接受）}}{\text{接受人数}}$$

P_{cT}为总命中率，P_{cp}为正命中率。

此种方法不大适用于高考效度的研究，而适用于小团体的职业选拔。

影响预测效度的因素有个体心理特性的同质性，效标的可靠，考试的长度，考试的难度及分数合成方法等许多方面，这需要深入进行研究。总之，在估计效度时，无论哪一种类，哪种计算法，都需要有目的地选用，方能有效。

考试的效度系数一般在0.40到0.70之间。美国大学入学考试效度要求也就如此。效度系数太低的考试对于考试的目标来论，是没有实际意义的。

三、提高效度的途径

提高考试的效度，旨在使考试的结果与考试的目标有更高的相关性，从而更好地实现考试的目的。为此，应提高考试的目标与考试功能，考试方法与考试内容，考试内容与考试目标等三对相关性上，以及考试过程的客观性，成绩使用的合理性等方面寻找提高效度的途径。

第一，合理地规定考试目标，使考试目标与考试功能有更高的相关性。

考试目标与考试功能相关与否，决定、制约着考试的效度。它们的高度相关是考试效度高的前提和条件。其理由是考试能够实现既定的目标，达到一定的要求，归根到底是因为考试具有与此目标相关的功能，所以，为提高考试的效度，

在规定考试目标时，应尽可能使之与考试所能发挥的功能相一致。

第二，科学地确定考试内容，使考试内容与考试目标有更高的相关性。

决定考试目标是重要的，科学地确定考试内容也是重要的，更重要的是使考试的目标与考试的内容有更高的相关性。规定的考试目标，是通过确定的考试内容来反映的。规定的考试目标与考试所具有的功能相关性再高，考试内容确定得不合适、不科学，而严密的考试也无法准确地反映考试目标。考试目标与内容的关系是辩证统一的。但是要把握好它们之间的高度相关程度是困难的。为此，为提高考试之效度，必须科学地确定考试内容，使所要考的内容恰恰是最能反映考试目标的内容。

第三，选择恰当的考试方法，使考试的内容与形式统一。

确定考试内容是解决考什么的问题，怎样考则是一个方法问题，是考的形式问题，内容与形式的问题就是内容决定形式，形式反作用于内容。一定考试的内容决定该考试的形式。一定考试的形式反映一定的考试内容。决定了形式的内容，考试设计者的首要任务是选择最适宜的考试方法，恰当地解决怎么考的问题，使实际考的恰恰是所要考的问题。解决好形式和方法。是提高考试效度的重要方面。

第四，合理地使用考试成绩

考试的成绩并不等于一系列分数，它究竟说明了什么？仅仅能说明什么？如何使用才能更好地实现目标，在有些考试中并不是一目了然的。为了提高考试的效度，提高考试成绩的使用效果，必须进行探究，更合理地使用它。这将在下

章讨论。

第五，加强考试的组织管理，提高考试的客观性、使考试有较高效度。

第三节　考试的难度与区分度

一、难度的概念与计算方法

考试的难度是表征考生解答试卷的难易程度的指标。通常用通过率来刻划。即：

$$通过率(P)=\frac{答对人数(R)}{总人数(N)}$$

（对试题可按以实得分数与满分之比表示）

通过率为难度不是十分妥当。因而用不通过率＋表出难度，即：　$f=1-P$

第j题的难度f_j可用下式计算：

$$f_j=1-\frac{\overline{x_j}}{a_j}$$

式中a_j是j题的满分，$\overline{x_j}=\frac{1}{n}\sum_{i=1}^{n}x_{ij}$　$j=1,2,\cdots\cdots,k$,

X_{ij}表是第i号学生，第j道题目的得分，$\overline{x_j}$是考生在第j道题上所得分数的平均值，当$\overline{x_j}=0$，第j试题难度$f_j=1$，表示这题难度最大；当$\overline{x_j}=a_j$时，则$f_j=0$。即没有任何难度。

试题的难度，并不完全是由试题本身的复杂程度决定的，它是一个相对量。它还与考生对该题的适应程度有关。我们

常常会遇到这种情况：有的题本身很简单，但考生大部分没有准备。结果得分低，通过率也低，有的题本身较得复杂，但考生大部分准备好了，结果得分高，通过率也高。实际上，试题难度所反映的是特定一组考生对该题作业的困难程度。

举例说明：

例：假设某班有50人，试题1、2、3、4均为多重选择题，每题有A、B、C、D四个答案，选择各答案的人数如下表：

假设的考试结果N＝50人

题号	选择答案个数			
	A	B	C	D
1	10*	20	15	5
2	25*	4	6	15
3	15	7	29*	8
4	21	22	5	2*

注：*表示此答案为正确答案。

那么，四个题的通过率分别为：

$$P_1=\frac{10}{50}=0.2=20\%$$

$$P_2=\frac{25}{50}=0.5=50\%$$

$$P_3=\frac{20}{50}=0.4=40\%$$

$$P_4=\frac{2}{50}=0.04=4\%$$

假设以P＝35%为等难度，则第二题较易，三题适中，四题最难．根据上表的数据，四题的不通过率就分别是：

$f_1=1-20\%=80\%$

$f_2=1-50\%=50\%$

$f_3=1-40\%=60\%$

$f_4=1-4\%=96\%$

二、区分度的概念与计算方法

考试的区分度是表示考试区分能力大小的指标。考试的区分能力，就是考试在用于考试时使水平高的考生得高分，水平低的得低分的倾向力。考试的区分度又可以说是试题能够把不同水平的考生按程度高低区分开来的程度。

考试的区分度是反映考试题对学生水平鉴别能力的指标。是试题能够把不同水平的学生按程度高低区分开来的度量，是试题质量的指标之一。区分度高的试题，考生好的考生得分高，学习差的考生得低分；区分度高的试题，学习好的与差的考生区分不出来。得分高低不太正常。某一试题的区分度，就是一组考生在该题目的得分与这组考生真实分数的相关程度。由于考生的真实分数我们并不知道，多次考试求其平均分数值的方法也不宜采用，我们可用某次考试的该组考生的得分代替他们的真实分数，转而计算这组考生在该题的得分与某次考试中他们所得分数的相关系数。当然，最方便

的办法，就是求考生该题得分与考生在包括该题的试卷上所得总分的相关系数。其计算公式为：

$$D=\frac{\sum_{i=1}^{h}(x_i-\overline{x})(y_i-\overline{y})}{nS_xS_y}$$

式中：D——某题的区分度；

X_i——第i个考生在该题上的得分；

$\overline{x}$——n个考生对该题平均得分；

S_x——n个考生该题得分的标准差；

y_i——第i个考生的总分；

$\overline{y}$——n个考生的平均总分；

S_y——n个考生试卷总分的标准差。

标准差S_x、S_y的计算公式是：

$$S_x=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n}};$$

$$S_x=\sqrt{\frac{\sum_{i=1}^{n}(y_i-\overline{y})^2}{n}}$$

为了估计区分度，可将考生分成高分组和低分组，并分别求出二组的通过率P_h和P_l，然后再求出P_h和P_l的差异D即

$D=P_h-P_l$

很明显，如果$P_h>P_l$，则$D>0$，说明高分组的考生通过率高，题目具有区分度。可以说，D值越大，区分度越高，

若D<0，则表明题目有问题。总之，D可以反映题目得分高低与总分之间的关系。因而可以作为区分度的指标。

另外，我们还可以用题得分与总分间的点的相关r_b来作为区分度的指标。即，若令x_p的通过题目的考生总分的平均数；S_x为所有考生总分的标准差，y为正态分布中面积，P是对应点的纵线高度，则：

$$r_b=\frac{\overline{x_p}-\overline{x_t}}{S_x}\cdot\frac{P}{y} \quad 或 \quad r_b=\frac{\overline{x_p}-\overline{x_y}}{S_x}\cdot\sqrt{P\cdot Q}$$

r_b的数值范围为－1～＋1，r_b的数值越大，该题区分度越高，在比较好的考试中，要求$r_b \geqslant 0.3$，r_b的最佳数值$0.4 \leqslant r_b \leqslant 0.9$。　例：假设10名考生某次考试的总分及在第1题上的选择如下表所示：

考生序号	总分	在第1题上的选择＊	考生序号	总分	在第1题上的选择＊
1	70	B	6	40	C
2	70	C	7	60	D
3	60	D	8	75	B
4	30	C	9	90	C
5	55	B	10	100	B

＊　B为标准答案。

根据上表的数据，可以求出：$P=4/10=0.4$

$\overline{x_p}=(70+55+75+100)/4=75$

$\overline{x_t}=(70+60+70+30+55+40+60+75+90+100)/10$
$=65$

$$S_x=\sqrt{(5^2+5^2+(-5)^2+(-3.5)^2+(-10)^2+(-25)^2+(-5)^2+(-5)^2+10^2+25^2+35^2/10}$$
$=20$

据正态表查出与P＝0.4点相对应的纵线高度为：

$y=0.39$，　那么$r_b=\dfrac{\overline{x_p}-\overline{x_t}}{S_x}\cdot\dfrac{P}{y}$

$$=\frac{75-65}{30}\cdot\frac{0.4}{0.39}=0.51$$

在最佳数值范围内，说明该题具有良好的区分度。

在考生人数较多时，用上述方法人工计算求出r_b，工作量太大，则可用计算机按公式求出r_b，不仅省时、省力，且所得结果也更加准确、可靠。或者也可采用查表求r_b即将总分从高向低排。然后根据这两个通过率便可用弗拉南根方法直接查表得出的列相关系数的估计值r_b。

三、难度与区分度的关系

美国戴德刊创造了一个试题分析的方法，可使教师较快知道所出试题的难度，具体方法如下：

试卷批改后，把参加考试者按成绩排队，找出总人数正中间一个人的成绩作为分界线（若人数为偶数，那么将中间

二人看成一人），分界线以上的称为“高；”用英文“high”的字头“h”表示，分界线以下者为“低”，用英文“low”的字头“l”表示，那么“h”和“l”的人数各半，而后列表逐题进行检查：

表格设计如下：

题 序	h	l	h+l	h-l
1				
2				
3				
:				
:				

把每一题h组与l组答对人数分别填在h和l栏内，算出h+l和h-l

难　度：$f=1-(h+l)/n$

区分度：$r_b=(h-l)/n$，式中n为参加考试的人数。

例：假定某班50人参加考试，试卷改完后还给学生。做法首先找出成绩的中间点，即由高到低排列，居中那位学生的成绩，如第25名学生成绩。然后将学生分为两组：中间点成绩以上者，称为“高”组（h），中间点以下者称为“低”组（l）。

分析表为：

题目	h（人数）	l（人数）	h+l	(h+l)/n	h-l	(h-l)/n
1	25	25	50	100%	0	0
2	25	9	34	68%	16	32%
3	18	15	33	66%	3	6%
4	9	18	27	54%	-9	-18%
5	16	15	31	62%	1	2%
6	15	2	23	46%	7	14%
7	10	10	20	40%	0	0

上表看出题目的难度是$f_7>f_6>f_5>f_4>f_3>f_2>f_1$而第一，七题的区分度为0，可以表明，题目太难或太易，区分度小。第四题的区分度为负。说明试题有问题，高分组学生答对的比低分组的还少，此题出得不够好，要修改。第二题难度为68%，区分度为32%，说明此题难度适中，区分度较好，可以供以后参考。

综上所述，在题目分析中，区分度和难度虽是两个不同的概念，但二者存在一定的联系。假如难度过小或过大，则即使实际能力有高有低，但都可能通过或都不通过。这样题目就没有区分度了。一般说，调整试题的难度是提高试题的区分度的重要方法，难度适中的题目，往往有较高的区分度。

试题的区分度与试卷的信度有密切的关系。试题的区分度越高，它对提高试卷信度的贡献就越大，试题的质量就越好。一般认为0.70到0.40，试题很好；0.40至0.30，试题较好；0.30至0.20，试题较差，应予改进；小于0.20，试题很差，应予淘汰。

四、提高考试难度与区分度的途径

研究如何提高考试的难度与区分度，是考试学的任务之一。难度与区分度严重影响考试的信度和效度。决定学生的成绩分布。高质量的考试，必须对提高考试的难度与区分度。

第一，正确认识难度与区分度在考试中的作用。难度和区分度是两次评价试题的重要指标，加上效度和信度，就能比较客观地反映试题的质量。反映考生的不同知识水平和智力差异。也就是说，评价考试质量的优劣，并不只看考试成绩的高低.事实上如果成绩普遍高，往往考试的质量反而很低，因而这样的考试结果，实质上是否定了考生之间的差异。比如，目前有的高校，仍然是由任课教师本人进行复习，命题和评卷，最后成绩的高低与教师的复习的方法，试题的难度，评卷的标准等都有直接关系。十分清楚，这样的考试，即使成绩很高，也不能说明教师的教学质量就高。为此，明确难度与区分度在考试中的重要作用是提高考试质量的思想基础。

第二，保证试题要有适当难度。

成功的考试首先保证试题的难易程度符合被考对象的实际情况，如前所述，难度对区分度的影响很大，只有保证试题都有一定的难度，考试的区分度才能提高。同时，还要认识到，如果考试的整个难度对各个试题作出的分配不同，区分度也就不同，即如果把试题的难度过于集中，则区分等级就要减少，所具有的鉴别力自然降低；如果把难度适当分散于许多试题中，区分的等级就会增加，所具有的鉴别力自然提高。为此，提高区分度，应把难度的分配做到相对平均并

有所侧重，以此通过增加鉴别的等级来达到高区分度的目的。

第三，增加试题的类型和条目。

试题的取样深，长度短，涉及教材知识面窄，这样的试卷，要准确考查出考生的掌握知识水平和智力的差异就很难办到。考生凭猜测和重点复习就能猜中考题的大部分内容，考试结果，水平相差不显著，区分度必然降低，因此要提高区分度，必须从考题的绝对数量上有所增加，通过增加试题数量，也有利于提高信度，而信度的提高，又反过来提高难度和区分度。

第四，加强对考试质量的评价工作。

搞好对考试成绩的统计分析和对考试工作的科学评价，对考生的真实情况有重要意义。考试的难度与区分度与考试方法和管理方法有密切的关系。如果考试的方法和管理方法是缺乏先进性、准确性和科学性的，常常只注意对考试成绩作一些经验性的描述，而缺少对数值资料的统计分析，那么考试难度与区分度就会受到严重的影响。为了提高难度、区分度，必须加强对考试质量的评价工作，这是提高考试难度、区分度的根本途径。

第四章　命题总述

第一节　命题计划的编制

一、编制命题计划的目的

命题计划是试题如何编制，试卷如何组成的计划，是供命题者编制试题和试卷的依据。目前，由于试题库建立不完善，一般考试都需要命题。因而都需要编制命题计划，编制命题计划是考试设计的重要工作之一。

命题是考试工作的核心环节，要提高考试工作的质量，关键是提高命题的科学性。命题工作的科学性主要体现在代表性与针对性。代表性是指试题取样能足够反映考试内容，针对性是指试题编制本身要合理。对不同的考试对象能有不同的体现。考试这种抽样测量。命题计划所确定的样本对于总体的分布和抽取的具体原则，对于命题的科学性，从而对整个考试工作的质量有很大的影响。

仔细研究过去的考试就会发现，许多重要考试的命题是不够理想的，许多命题不理想不仅与命题人员专业知识缺乏和命题经验不足有关，而且与命题计划不周，试题取样代表性不够紧密相关。1979年高考试题就是一个典型的例子。强调命题计划编制的重要性是很有必要的。

二、命题计划的编制

任何事物都有一定的规律。坚持一定的原则。编制命题计划，必须遵循如下原则。

（一）符合考试标准。

命题计划中的各项原则要求，特别是命题的分布范围、难易深浅、考查重点等，必须完全符合考试大纲，既不能扩大范围，又不能缩小范围，既不能提高要求，又不能降低标准。考试标准是关于考试范围、考试内容及各部分要求掌握的程度的规定。它是考生备考和教师命题的基本依据之一。考试标准是根据考试目标制定的。坚持考试标准是实现考试目标的要求。

（二）计划编出的考卷要充分反映考试内容。使其内容有足够的代表性。

试题数量的多少与考试衡量指标有一定的关系，要提高考试指标绝不是将试题的数量无限增多。况且，考生答卷的时间等因素的限制，考题不可能过多。在试题数量确定的条件下，试卷对考试内容应有足够的代表性。因此，命题计划对于试卷组成的质的规定，必须符合考试大纲。对于试卷组成的量的规定，应能充分反映考试内容。这是一条重要的原则，只有坚持这条原则，才能减少取样不足所带来测试误差。

（三）掌握试题的数量和占分的比例，按学科知识的部分和心理能力的层次进行抽样。

众所周知，考试内容所要考查的是总体，而试题则是我们从考查内容总体中抽取的样本。要使样本能够代表总体，亦即考生对试题的解答情况能够代表他们对于整个考试内容

的掌握情况，必须坚持随机抽样的方法。即是考试的内容都有同等机会被考到，所有被考查的内容的概率相同。要想提高样本的代表性，就得将总体按某种特征进行分割，以不同的类型和不同的层次形成网状结构，根据各类型和部分所含总体成分的多少，进行每类或部分中的抽样比例再随机抽取，即按分层随机抽样。

命题计划的内容是非常丰富的，它不仅有对试题和试卷编制的要求，具体说明考试的目标和内容范围，考试方法和试题类型，编制试题和组合试卷的要求等的“命题大纲”，而且还有试卷中试题的分布规定，具体规定出考试内容中各部分的试题数量和占分比例的“双向方格表”。

三、命题大纲举例

全国高等教育自学考试指导委员会编制的哲学命题试行大纲（摘要）（摘自《哲学命题试行大纲及试测题》，辽宁人民出版社）。

（一）命题工作的指导思想

马克思主义哲学是高等教育自学考试各专业（除哲学专业外）的公共必考课，这次哲学命题，除为此次统一考试提供试卷外，并为建立题库打下初步基础。

命题是考试的核心工作，这次命题，要在各地命题经验的基础上提高一步。

这次哲学命题，是第一次全国统一命题，是高等教育自学考试工作中采取的一项重要措施，必须慎重从事，严肃对待。通过这次命题，应为今后其他课程的全国统一命题和大范围协作命题探索经验。

高等教育自学考试是考核自学应考者高等教育学历的国家考试。对于一门课程来说，务必使考试及格者确实达到全日制普通高等学校相应课程的结业水平。

编制的试题和组成的试卷，通过考试要能够正确引导个人自学和社会助学，树立良好的学风。要引导他们认真全面地学习教材，掌握系统知识，培养、提高分析问题和解决问题的能力。

要编制客观性试题，增加客观性试题在试卷中的比重，为高等教育自学考试采用客观性试题探索经验。

（二）命题的原则

1. 坚持考试标准，掌握好及格线。哲学课的考试标准，就是普通高等学校公共哲学课的结业水平。高等教育自学考试《马克思主义哲学原理自学考试大纲》是这次命题标准的依据。

高等教育自学考试是考核应试者是否达到规定的标准，是通过考试成绩是否及格来衡量的。因此，考试标准集中反映在成绩及格线（60分）的确定上，编制试题和组成试卷，必须着重掌握好及格线，成绩在及格线以上的应考者，应确实达到哲学自学考试大纲的基本要求。

2. 正确掌握命题范围，按照指定教材的内容命题。不要扩大或缩小命题的范围，也不要提高或降低考查的深度。

3. 适当增加题量，扩大覆盖面。全体命题教师编制的试题总体应能覆盖教材各部分具有考查意义的内容。

4. 增加客观性试题，主观性试题要考虑便于阅卷评分。客观性试题有多种类型，这次哲学命题只采用“单项选择题”和“多项选择题”。

5. 难易适度，并有恰当的层次。

总之，命题的原则可以有若干条，关键是要正确掌握考试标准，结合自学考试的特点，扩大试题覆盖面，使凡是认真，全面和系统地学习了指定教材，并能运用基本原理分析，解决一定的实际问题，达到了大纲要求的，应当能够及格。

（三）命题的具体要求

1. 编制试题要覆盖到每个章节和重点项目，对于同一内容，可以从不同角度提出问题，编制试题。

2. 试题分布，按照教材的份量来安排，绪论和辨证唯物主义部分约占百分之六十，历史唯物主义部分约占百分之四十。辨证唯物主义的三个部分（即唯物主义、辨证法、认识论）原则上按各占百分之二十掌握。

3. 每份试卷约包含50－70个试题。其中客观性试题约占50－60个，每题1分。

四、认知目标分类命题

根据教育目标认知范畴分类系统来拟制试题，可保证试卷不致编重于考查某一类的认知活动。试卷编制，第一个步骤就是确定该试卷达到某一种功能。继后决定各题的考查目标，跟着就是试卷蓝图的编订。

美国教育家布卢姆（Bloom）曾将认知活动的教育目标从低级到高级划分为六类：

①认知：指回忆或认知事实、规则或概念的能力。认知具体事物，处理具体事情的方法和有关方面的一般概念和抽象概念等。

②理解：指理解事实和概念的能力。如翻译、阐释、推

断等。

③应用：指利用事实和概念解决新问题的能力。如规则、方法、概念的应用等。

④分析：指辨别整体中的各个局部并认识其相互联系的能力。如成分、关系、组织原理的分析等。

⑤综合：指把有关局部综合成新的整体的能力。如交流、计划、从中抽象出若干种关系等。

⑥评价：指判断，比较不同方法、结果等的能力。如使用内部证据、外部标准等。

布卢姆的关于教育目标的上述分类方法，已为多数教育专家所接受，许多考试的双向方格表，能力的"一项"就是按此方法分类的。

例1. 大学水平的教育测量理解程度测验计划书（罗伯特·L·艾伯尔著，漆书青等译：《教育测量纲要》，第74页）。

（1）项目的形式	数目
多项选择法	50
（2）作业的种类	项目
专门名词	5
实际的知识	10
概括	10
解释	10
推算	5
预言	5
推荐活动	5
合计	50

（3）内容的范围

教育测量的本质	2
教育测量的历史	2
统计技术	7
发现和选择测验	3
测验与教育目的	3
教育自编测验	4
测验的试测分析	2
初等学校的测验	5
中等学校的测验	4
教育性向	5
个性与调节	2
观察的技术	2
学校的测验计划	5
测量结果的运用	4
合计	50

（4）项目的难度

预期平均错误的百分比	30%
错误百分比的全距	10—60%

例2："力学"测验蓝图

	记忆	理解	应用	综合	项目数	百分比
"力"概念的演变	20				20	11.70
"力"的类别	20	20			40	23.34
二维力	15			25	40	23.34
三维力	10			25	35	20.46
物质的相互作用	8		8	20	36	21.16

例3：英语水平考试（EPT）的双向方格表（摘自桂诗春教授在北京国际标准化考试讨论会上的发言材料。）

	知识	理解	应用	分析	综合	项目数	百分比	时间
语法	10		10			20	12.5	20分
词汇	20					20	12.5	60分
阅读		40				40	25	
综合填充				20		20	12.5	20分
听力		35				35	21.9	30分
写作					25	25	15.6	30分
总计	30	75	10	20	25	165	100	160分

任何分类都有其弊和利，试题按所考核的具体心理能力进行划分，常常受到界限不易划清的限制，一些试题很难按心理能力明确归类，为此，有人认为应将这种能力项目改为“基础题、综合题、提高题”三类，“基础题”包含“知识”、“理解”、“应用”三个层次要求。“综合题”包含“分析”、“综合”二个层次的要求。“提高题”包含“综合”、“评价”二个层次的要求。试题的比例为：7：3：1。

分类的基本出发点，应该根据考试的目的、任务和考试对象的实际情况为基准的。只有根据外在特征分类和成功解答问题所需要的心理能力分类，才能提高考试的命题质量。

第二节　命题的基本原则

一、命题工作的重要性

从系统的观点看，命题是考试工作重要因素，核心环节。

它是考试目标、考试内容与考试方法，编制考试计划到印题，考试实施，评卷和成绩的解释与使用的中介与桥梁。没有命题，就没有考试。在不同的程度上，命题就是按照命题计划编制题和试卷，为考试制造测试工具。

我们可以想象，尽管考试设计得再合理，计划编制得再完善，如果命题工作没做好，那么也不会产生好的考试效果。同样，考试过程再严密，数据处理再科学，命题工作没有做好，试卷的质量很低，也同样不可能很好地实现考试目标。

考试试题的作用和价值并不随考试的过去而消失。它常常作为一定时期，一定阶段的教育测量水平的反映，是重要的历史资料。优秀的试题往往会被互相传抄，甚至记录成册，不少习题集和习题集解答对人才的培养和发现，产生了深远的影响。

提高考试质量，试题质量是关键。试题质量又依赖于命题工作。因此，做好命题工作，提高试题质量对考试来说有更为深广的意义。

二、命题的基本原则

命题是考试的中心环节，考试的指导思想也主要体现在命题上。它是一个严肃的问题。任何图简单省事，命题过易或过难，出偏题，都是有害无益的，既不能真实反映学习水平，也不利于鼓励考生上进。因此，必须十分认真严肃地对待考试命题，提高命题质量，使考试能充分发挥其功能和作用。

命题的基本原则，主要体现在以下几方面：

（一）命题主要依据命题大纲，必须全面反映大纲的广

度和深度。

命题的依据是命题大纲，既要有反映考生掌握基本知识、基本理论和基本技能的题目，又需要考核考生灵活运用所学的理论去分析和解决问题的能力。要使命题全面反映大纲的广度和深度，考题数量要足够，复盖面要大，重视“三基”、“能力”的考核。命题依据命题大纲是命题的第一原则。

（二）命题要有利考核和促进考生提高智能。

命题要体现考知识又考能力的要求，贯彻知识积累与智力、能力发展相结合的基本思想。考试不仅考核考生知识掌握如何，而且通过考试，促进智能的发展和提高。要使试题类型多样化，从不同侧面考核考生的知识与智能。在组合试卷时；要考虑试题多种类型与功能。选择最佳组合结构，这是命题的又一基本原则。

（三）命题要讲求层次，要有难度台阶，才能在评分上拉开差距，提高分数的置信度。

试题如果太容易，不能激发考生的积极性，考不出真实水平。如果试题太难，超出考生的实际程度与能力，易使考生失去信心，同样考不出考生的真实水平，太难太易都不符合命题要求。命题的难度，应从大多数中等水平考生出发，形成难度台阶，使大多数考生能在解答试题上形成差距。形成一定难易距离和梯度，才能考出考生不同的水平；以便通过考试把优秀生与一般学生区别开来。当然要解决这一问题，绝非难度与区分度的问题。对信度，效度也同样要研究。要达到理想程度，命题应当有较高的信度、效度、区分度和一定的难度。这是命题的重要原则。

（四）命题要注重发挥考题对考生学习方法的引导作用。

众所周知，考生一般说来重视考试，自觉和不自觉地受到考试“指挥棒”的影响。当然，不能赞成把考生引向为分而学习，学习专门为了应付考试。但是，要充分估计考试对考生的引导作用。通过命题来引导考生掌握正确的学习方法：考试象磁石一样，吸引考生朝哪一方面下功夫。如果考题偏重记忆，考生就要去死记硬背；如果考题着重灵活应用，学生就要开动脑筋，思考钻研问题。

命题是一项复杂的智力劳动，编出一个好的试题、一份好的试卷，要求命题者除了完全掌握考核学科的知识外，还要有一定的命题技巧。一般说来，一份好的试卷，除了题目本身的科学、合理外，还应该具备这样几个条件。

1. 提出的问题，设置的解题任务，是考试内容中实质性的东西。

2. 问题的正确答案，是有定论的。但最好不是教科书上的原话。

3. 提出的问题的方式，设置的解题任务的情境，是新颖的，不落俗套的。

4. 问题的含意是明确的，而不是暧昧不清或模棱两可的；用语是简练的、准确的、而不是罗嗦的，费解的；解答的要求是清楚的，具体的、而不是模糊的，可随意理解的。

三、命题的任务

命题工作首先是研究考试大纲和命题计划，明确考试的目的，考试的性质，考试的对象，考查哪些知识和智能，试题的形式和数目，试题数量和分数在各部分中的分配比例，考查的重点，及其他要求。其次根据命题计划编制试题，同

时给出每一试题的答案。编制试题的数量至少是需要的二倍以上。并对编出的试题逐道审查，修改和筛选，使通过备用的试题及其答案都科学、合理、用语准确。同时注明各备用题的预计难度，考查部分的能力层次。最后制订评分标准。

总起来讲，命题的任务就是按命题计划编制试卷和试卷使用的"说明书"。即命题工作的三要素：一是编制至少两份具有相同等效力的试卷，一份为正卷，一份为副卷；二是编写试卷的参考答案；三是制订试卷的评分标准。

第三节　预试概述

一、预试的意义

衡量考试质量的指标无一不与考生的状况相关。考生的特点，是试题编制过程必须认真考虑的重要因素。考试所用试卷的质量不仅决定于试卷反映的内容，符合考试大纲的程度，而且还取决于反映考生特点的程度。考前进行预试，选取一定数量，对考生有一定代表性的人员，按正式考试的要求，解答准备使用的试卷或部分试题，借以研究试卷的质量，并据以修改、调查试卷。听取受试者的反映，分析试测的结果。对于提高试卷考查的针对性、防止意外情况的发生，是十分必要性。

二、预试的组织要求

预试的成败，对正式考试成功与否影响很大，有时起着决定性作用。因此，对试测的组织要进行认真的研究。提出

一定的要求。

首先，预试对象对正式考生要有一定的代表性。

预试的目的就是要提高试卷考查的针对性，防止意外情况发生。如果选取的预试对象与考试对象相差很大，那么，预试所提供的信息，是无用的。因此，选取预试对象要有一定的代表性。他们可是考生的一部分，或是刚刚参加过同类考试的人员，或是具有某种可比性的其他人员。对于规模较大的社会性考试，还要考虑不同地区，不同职别，不同年龄、不同水平的抽样比例。代表性越大越好，人数越多越好。

其次，要做好预试对象的思想工作，使预试具有真正的价值。

预试的组织和实施，应按正式考试的要求进行。由于预试毕竟不同于正式考试，受试者特别是不参加正式考试的那部分受试者，往往缺乏一定程度的考虑和积极应考的心理。因此，应特别做好受试者的思想工作，使之理解预试的意义。采取认真严肃的态度，自觉地进行配合。必要时，还可采取一些促使受试者努力作答，争取获得好成绩的鼓励性措施。此外，还要向受试者说明考试的内容和要求，并给予较充分的复习备考的时间。

再次，做好考试前的试题保密工作。

预试要求受试人员尽可能多一些，而受试人数越多，试题越容易扩散，造成漏题。因此必须认真做好考试前的保密工作。以保证考试的顺利成功。

三、对预试结果的分析

预试的目的是提供关于考生对试题适应情况的信息。但

预试的反馈材料（试卷和记录材料）还只是一种初级信息，只有经过科学处理后才能转换为直接意义的信息。这个处理的过程就是对预试结果的统计分析。

统计分析的项目有：试题的难度、区分度、试卷和考试的信度和效度，选择题备选答案的误道分析。

此外，根据修改试卷的具体需要和预试实施过程的记录内容，还可以进行考生心理分析，解答题时间的统计分析等。

除对预试材料进行统计分析外，还可通过分别召开受试者座谈会，评卷教师座谈会等方式收集对试题的反映，进行定性分析。

第四节　题库的建立

一、题库概述

所谓题库，就是指经过一系列的分析鉴定后，符合一定质量标准的试题的集合。一个学科的某类考试的题库，就是能提供这个学科该类考试所用试题的试题总体。多年来传统的试题来源，无非是自拟定，手头保存的，或者从别人那里选来的（包括从各种习题集中选来的）。其中除了临时自拟之外，其余的来源实际上都可以看做是一种题库，只不过以前没有这种提法罢了。随着考试方法与理论的发展，使得考试规模扩大，次数频繁，人数增加。在这种情况下，为了从试题的质量上来保证考试成绩的可靠性，在试题的选择上就不能继续依靠上述的老方法了。这就很自然地产生了一个必须尽早建立题库的问题，以使各种类型的考试需要的大量的

高质量试题得以满足。

然而，一个完善的题库，要具备什么条件呢？哪些条件是必要的？

题库应备有考试大纲，教材，命题大纲或命题计划书及考试命题文件和题库组成，题库使用情况的详尽材料。同时必须具备如下条件：

1. 所储试题的总体能够覆盖考试大纲要求掌握的这个学科的全部内容，有多角度考查意义的内容都备有从不同角度考查的试题。

2. 每一试题都科学、合理、用语准确，并备有答案（或答案要点）和评分规定（或说明）。

3. 备有符合命题大纲要求的若干部分（如十至十二份）标准样卷。

4. 每一试题都制成试题卡片，卡片除试题及其答案外，还须注明考试的内容（章节）、题型、预计难度、满分量和评分规定及试题编制，审定人等项目。试题卡片须编号分类保存。

二、建立题库的意义

国外的考试委员会都很重视建立自己的题库。美国任何考试委员会都已经建立了题库，而且都贮存十几万道试题。建立题库的意义概括起来有以下几个方面：

1. 保证考试成绩的可靠性

为了保证考试成绩的可靠性，一是增加题量（对多项选题来说，一次考试不能少于30题试题），二是试题的难度和区分度要适中。这两方面的要求都可以从题库中得到满足。

因为题库中的试题一则数量大，二则每道题都是经过质量分析后注有难度指数的区分度指数的，有很大的挑选余地。

2. 提高考与学的效果

建立题库的过程，也促进了主考者对考试大纲和内容的理解。考生选作题库中的试题，也可以检查自己对所学的内容掌握的程度如何，从而使考生的学习要求向大纲要求的方向靠拢，提高了考与学的效果。

3. 保证考试科学化

各种类型的考试，诸如升学考试，阶段考试，毕业考试和业务统考等，都有其各自的考试大纲要求。在大纲中对不同章节，不同门类的要求比例是不同的，测验知识的测重面也是不同的。有了题库，读者就可以根据这些不同要求从题库中选择所需要的各种类型的试题，使得考试科学化。

三、如何建立题库

建立题库的本质就是收集试题。被收集的试题应该是经过一定的试题标准衡量，满足质量要求的题目。一道试题能否收入题库，一般应经过标准化处理和质量分析的过程，才能决定取舍，并且还应根据科学的发展和教材内容的更新而不断地加以充实，使试题始终保持其先进性。使题库充满着旺盛的生命力。

1. 试题的来源

题库中试题的来源。一方面是靠收集和征集试题，放开眼界。古今中外历次考试，各种书本中的例题和练习题，尽量多翻阅，多收集，多积累。采取“约稿”方式，请有关人员，按指定的要求分类出题。对于收集、征集和组织教师命

出的试题，要聘请有关学者和考试专家进行审定。另一方面是聘请专家集中命题。特别是有针对性的考试。

2. 试题的标准化处理

目前，各类型考试中，选择题已经越来越广泛地运用。但由于大多数人对于多种选择题的拟出还缺乏经验，加之各自的写作能力不同，对教材的理解程度上存在着差别，所以试题在拟出以后，还要经过集体讨论，经过标准化处理后，才能正式作为试题应用。这种根据一定的标准进行审定的过程就称为试题的标准化处理。

一道试题能否达到满意的标准。它必须满足一定要求：

（1）试题要涉及到本学科的一个或几个重要方面。要与应试者掌握的知识水平相适应。题目难度适当。

（2）对主要问题要作出明确的正确的陈述。用含混的意义不明确的字眼来陈述有可能使应试者产生误解。破坏试题的可靠性、为应试者作出解答提供合理依据所必需的限制条件等等，都是不符合标准化的。

（3）同正确答案无关的暗示是否已经除去，常见的暗示是：叙述上正确的答案比误道答案更长些，清楚些；用词上正确答案使用专业术语，干扰答案不仅不用专业术语，反而用一些不应使用的字眼等。造成无意中泄漏答案。

（4）为了使试题明确和完整，应该使用尽可能少的词句进行叙述。因为不必要的文字会增加应试者的阅读时间，应针对题目的内容和要求，选择适当的题型。

3. 试题的质量分析

试题经过上述的标准化处理后，虽然在正式考试中被采用，但这并没有完成该题被收入题库前的过程，还必须在考

试以后，再次根据一定的质量指标对其进行质量分析。衡量试题质量的常用指标有两项，即难度和区分度。此外，还应该同时分析无效答案的出现率，才能对试题是否收入题库做出最后的判断。

难度f：　$0.2 < f \leqslant 1$

区分度D：　$0.3 < D \leqslant 1$

另外对f、D还必须结合起来分析，以决定对试题的取舍。比如，f是低值本应放弃，但如果其D值却是满意的，而试题又没有找出内容上和用词上的错误，那么这道试题就应该保留入库，因为它能有效地对好的和非常好的应试者作出区分。

四、题库充实与更新

现代科学的迅速发展，必然促进教材内容的不断更新，不断加深。建立题库的工作也就增加了新的内容，那就是必须根据变化的情况不断充实新试题。题库的建立，绝不是工作的结果，还必须通过收集，征集试题，组织教师命题，审查修改等办法，特别是预试和分析每次考试结果的办法，不断充实，调整，完善和提高。

第五章　主观性试题命题

第一节　主观性试题的概述

一、主观性试题的概念

任何考试，都应该是客观的，应该是反映考生之间客观存在着的差异。反应在分数上，就是不同水平的考生得不同的分数。如果要使某种试题考得的分数完全是主观的，不反映考生的真实水平，那么，这种就不叫考试，不符合考试的基本要求。但是，如果要说考试完全是客观的，这也是不可能的。因为任何考试都不得不带有一些主观的色彩，不可能完全避免主考人主观因素的影响。为此，主观性考题究竟是什么？它的基本判别方法是什么？

考试的发展，使得考试题各式各样，比如填空题、填图题、改错题、判断题、概念题、列举题、计算题、选择题、证明题、论述题、作文题等等。要判断是否是主观性试题，是有一些困难的，有的试题就不能归为主观性试题。因此，这里谈“主观”是一个相对的字眼。就试题的正确答案看，主观性试题的正确答案可用多种方式表述，评卷员须凭主观经验给分。当然，考生在解答主观性试题时，则可以“自由应答”。实际上，主观性试题只是给出问题，要求考生写出问题的完整答案，根据考生所写答案的正确，完整程度给适

的分数。正是这评分上的“适当给分”，使得主观性试题的评分含有阅卷人主观因素的原因。

主观性试题，是出现得最早，历史最悠久，现在仍广泛采用的一种试题。它不仅是一种试题类型，而是具有某种特点的众多类型试题的总称。它的特点有三：第一，每题有超过一个可能正确的答案；第二，考题容许答题者自由发挥；第三，需要评卷者判断答案。美国的韦特曼在一九三三年曾经归纳了十一类问题，作为主观性试题。它们是：（1）什么事，什么人，什么时候，哪一个，哪个地方等；（2）列举；（3）概述；（4）描写；（5）对比；（6）比较；（7）说明；（8）讨论；（9）发挥；（10）总结；（11）评价。实际上远不止这些，比如作文就是一种典型的主观性试题。

主观性考试又称“论文式考试”或“旧法考试”。它的历史可追溯到两千多年前的中国。可以说主观性考试，始于并盛行于中国。西周国学中的大学订有年定期考查制度。据《礼记·学记》中记载，学生每年入学，每隔一年，必须考查他们的学业成绩。从隋唐开始，直到清代末年（1903年）“癸卯学制”颁布长达一千三百多年的中国封建科举制度被废除为止，我国都以科举的方法来选拔人才，这种方法叫做“开科取士”。唐代科举考试比较完备。唐的考试制度，对宋、元、明、清各代的科举考试也有很大的影响。后来各代的许多考试方法，追根溯源，多数发端于唐代。唐代科举考试方法有口试、帖经、墨义、策问、诗赋五种。宋朝科举考试的程序有“秋试”，“省试”和“殿试”。在欧美国家、旧法考试的产生比我国要晚得多，直到十九世纪初，这些国家对学生的成绩考核仍采有口试的方法。到了1837年，美国

教育家霍瑞斯曼才主张改革成绩评定方法。1845年，在他的领导下，波士顿学校委员会首先采用了书面考试的方法。

二、主观性试题的优缺点

主观性试题，历史悠久，经久而不衰，至今仍是考试中普遍采用的试题，它必然有许多独到的长处，优点。否则，它便没有存在的立足之地，将被其它型式的考试所代替。当然它也有自身的短处、缺点，不然的话，为什么其它的试题有生存，立足之地呢?

主观性试题的优点，主要体现在如下几方面：

第一，主观性试题能够在一定程度上反映考生解答问题的思维过程。

考生解答主观性试题，需用文字（书面语言）表述自己的见解。解答的文字，恰是思维过程的外在表现，考生答对与否，他是怎么思考的，科学否，合理否，新颖否，简捷否。答错了，错在什么地方，是考生根本不懂，还是一步失误，等等一系列问题。主观性试题能在一定的程度上反映出来。特别是对于诊断和反馈性考试有极其重要的意义。

第二、主观性试题能使考生较充分地表述自己的见解，展现自己的才华，提高考查的深度。

主观性试题，让考生自己编写问题的完整答案，这就可较好地展现考生的才华。主观性试题能做到让考生不只是复述一个概念，一个原理，一个事实；而且还能让考生表达自己的见解，如分析一种现象，评价一个事件，说明或论证一个问题。因此，这样的考试就能达到一定的深度。例如，主观性试题“从事物发展是前进性和曲折性的统一来阐述树立

共产主义理想的必要性和迫切性”，就要求考生作答时，就比只把这个问题所考核的内容，分解为许多要点，再提出几种可能的说法，仅让考生去作出简单的表态，在反映考生思想水平和分析问题能力的差异上，考查的深度要深得多。

第三、主观性试题能对具体知识进行综合考查。

主观性试题本身不给出答案。对问题的解答者——考生提供了回答问题的广阔天地。这样就比给出几个答案来供考生选择更有利发挥考生的真实水平。它不仅可以从小处着手，对部分知识进行考查，还可以从大处着眼，对知识，能力进行综合考查。比如：命题作文，它就可以考查考生独立构思写作的能力，即识字写字，用词造句，布局谋篇，运用语言文字表达思想感情的知识的综合应用。同时，它又可把写文章的能力分解为许多要点，分别考核选材的能力，划分层次能力，论证能力，驾驭和运用文字的能力，等等。

第四，主观性试题是容易实行，它照例问题很少，不论是油印或写在黑板上，考生都能看见。出题的困难也较少，不需要许多设备去构成一场考试，大多数学科都可以采用它。知识专家可自行出题，一般不需要试题编制专家的参考，应试者凭机遇在考场上猜题得分的可能性甚少。

主观性试题的缺点主要体现在如下几方面：

第一，论文式考试取样太小，覆盖面小、缺乏代表性，不能对考生作出全面考查，且易使学生解答一道主观性试题，少者写几句话，多者写一篇文章。少者用几分钟，多者用几十甚至上百分钟。因此，每次考试，只能出几道或十几道主观性试题，如果所考的是一门课程，许多章节甚至摊不上一道试题。虽然每道题都可能考到一定的深度，但考查面过小，

试题取样对于待考查的内容总体上往往缺乏足够的代表性。拿1981年全国高校统一招生考试的数学试题（理工农医类）和《全日制十年制学校中学教学大纲》（试行草案）规定的教学内容作比较，就可发现，整个中学数学内容多达39项，而1981年数学试卷加上附加题才有十道题，一些如初等函数，视图，统计初步，空间图形，微积分等主要内容都没有涉及，整个平面几何部分的试题也份量不足。这样就有可能侥幸得高分。使学生产生考试碰运气，投机取巧，猜测教师出题范围的心理。

第二，主观性试题的难易度、区分度不好掌握，主观性试题命题制卷后即可直接与考生见面，不经过预试，也不加分析和修正筛选，因而，凭经验命题很难与考生实际水平相符。笔者参加1984年高考阅卷时，从试卷和成绩统计数字上看出，命题意图是好的，题目也出得有水平，加强能力训练体现得较充分。但由于带有很大的主观性，结果难度都大大超过了考生实际水平，使高、低水平的成绩拉不开，造成选材困难。

第三，主观性试题评卷工作量太大，评分不客观，即使评阅几百份试卷也是一项非常繁重的工作。答卷的措词往往使人们不可能给予客观的评分。评卷员在两个不同的时间对同一答案的考卷不可能给予相同的评分。即是阅卷者的主观意志（如情绪、偏好等）及阅卷时的客观因素（如环境的条件，试卷出现的先后等）也都会左右分数。一个学生讲了这样一段话：“我希望我的历史教授在评定我的考卷之前没有同他的妻子争吵，而且吃了一顿美味的早餐”。他的话对主观性试题评分中常见的个人因素给予了一个生动的描述。这

样造成了评分带有很大的主观性。难以反映出考生的真实水平。著名教育家斯太奇（D·stach）将一份英文试卷请142名本科毕业的中学教师评分，得到35种评分结果，最低50分，最高90分，相差48分。为了批驳有人认为数学、物理、化学等评分比英文考试评分要客观一些，斯太奇将一份数学试卷请115位中学数学教师评分，得到41种评分结果，最低为28分，最高为92分，相差64分。国外教育界还有一件引为笑话之事，某年夏季，许多大学教授在评阅历史考卷，有一位教授为评阅方便起见，自己写了一份答案作为典范。不料这份范卷和其它待评考卷混在一起，给另一位教授评阅，竟得了一个不及格的分数。为慎重起见，其他教授对这份不及格的试卷重复评定，结果所得分数差别之大，竟从40分起直至98分为止。

第四，主观性试题评分对分数的解释不科学。主观性试题使用的是原始分数，由于它们不具有相同的单位和参照点，所以无法相加，也不能进行比较。一个分数，往往只能标志考生对试题解答的程度，而不能看出他在其所处的集体中居于什么地位。我们不能说语文科的85分会等价于数学科的85分。也不能因为去年期末数学考95分，今年期末考90分，就说今年的学习退步了。还有我们不能说甲、乙语文，数学都是80分，就断定他们的成绩一样好。原因很简单，即使是单位相同参照点不一样，也不能直接比较。

三、主观性试题的选用

众所周知，试题的选用都是要扬其长、避其短。我们分析题型，研究试题。不是企图说明哪一种绝对地优于另一种

（其实，只要能存在，就不可能没有优越性），而是研究在什么条件下，使用哪一种更有效，使得它的短处对于实际考试的有效性不构成实质上的干扰。

根据上述分析，主观性试题在以下几种情况下比较适宜。

1. 在着重考核考生综合运用能力的考试时。

2. 主要为了提供教学反馈信息和成就考试。

3. 考生数量较小的课堂考试。

4. 某一专项的高级考试。

5. 智力测验。

总之，主观性试题多应用于课堂的成绩考试，诊断考试和高级学习阶段的综合考试（如硕士、博士研究生的考试，大学生的某些综合性考试，博士后的考试）。主观性试题是大有用武之地的，它将在多种级别多种类型的考试中发挥巨大的作用。

为了讨论的方便，本章把主观性试题归成三大类，即简答题，论述题和作文题。下面将逐节加以论述。

第二节　简答题的编制

一、简答题的特点

简答题不是一种主观性试题，而是一类主观性试题。它是指答案比较简单的主观性试题，是主观性试题中的“小题”，它多半是考核考生的基础知识的理解和巩固的程度，反映考生在知识占有上的差异。它特别适于考核基本史实，基本概念、基本原理等。

简答题一般包括四种主观性试题。

（1）简释题，解释概念，名词等。如，社会主义的工资，政党、工会、无产阶级、货币、剩余价值、时间、运动、空间、矛盾、物质等。

（2）直接问答，回答“是什么”的问题，即韦特漫列举的“什么事，什么人，什么时候，哪一个”等。如象，社会主义国家奉行和平外交政策的理论依据是什么？中国共产党处理和发展同各国政党的关系的基本原则是什么？贝克莱说：“存在就是被感知”，此种哲学的观点及主要的错误是什么？

（3）列举题，举例说明问题。如象，举个例子说明什么叫做反馈，中国共产党领导的第一次工运高潮中的三次大罢工，列举空想社会主义的三位代表人物等等。

（4）扼要说明题，简要叙述题。如图示出固定资本和流动资本与不变资本和可变资本的区别与联系？垄断组织的主要形式是怎样变化的？分别判断下述说法的对错，并扼要说明理由：“级差地租等于商品价值与生产价格的差额”，“级差地租等于剩余价值与平均利润的差额”，“级差地租等于社会生产价格与个别生产价格的差额”，“级差地租等于超额利润与平均利润的差额”。

二、简答题的编制

简答题是主观性试题中的比较带客观性的试题，设计和编制中，要特别注意发挥其评分较为客观的长处，使之考核内容具体而不空泛；使考生作凝缩性的回答而不作扩展性的回答；使它的正确解答较为简短而规范。

简答题比较机动灵活，能从不同角度和方向发问。因此，在设计和编制中，要注意发挥其特点和长处，以全方位的角度和方向发问、提问，使编制试题最大限度地增大考核的准确性和深度。命题时既可以让考生直接回答某个概念和原理的内容，又可以让他们找出似是而非的说法中的错误之处，并给予更正。还可以让他们判断其对错并批驳它的谬误和论证它的正确性，等等。当然，编制设计是尽量从大处着眼，从小处着手。当考核掌握一个完整知识中容易出错而又十分重要的关键问题和知识体系中的构架或对主要特征的概括。

比如“一台机器，每年提取其一定比例的折旧基金，当年提取的折旧额和当年固定资本实物补偿的价值量是否一致？当全部折旧基金等于机器原值，而机器仍然还在使用，这时是否还有价值转移？”。

三、简答题的改善

简答题比较容易编制，人们也常常忽略对它的研究。要把答题出得“活”，考出深度，考出水平，而答案又简洁规范，易于阅卷给分，却不是一件易事。

由于人们对简答题的认识不足，轻视这类题的研究。编制过程常出一些缺点。归纳起来，主要反映在如下几方面：编制题目的形式单一，题目千篇一律，大都问“什么是”或“是什么”；因而考生思维的方式单一，都是回忆既定的现成答案；解答问题的方式单一，都是在问号下面的空格写答案，开头语都是所谓“××是指……”；设计题目的角度单一，个个都让考生正面回答；正确答案的出处往往是教科书

上的现成语句。使得考生的思维方式和学习方式都很刻板……。统而言之，就是题目单一、题目死板。

对简答题的改善，旨在充分发挥它自身的特点，避其短处。简答题是主观性试题中的“客观性试题”。因此，要扬其“客观性”之长处，把简答题客观化。使其答案更加标准，提高评卷的信度与效度。例如：精神文明文化建设方面的社会性质是指：（　　）

1. 阶级性，　2. 政治性，　3. 思想性，　4. 服务方面。

“灵活”是简答题的又一长处，它能够更方便地根据需要创设不同的问题情境，从不同角度进行考核，增加考查的深度。由于长期不重视简答题的研究使其本身的“灵活”性与试题的单一性产生矛盾，从形式和内涵来看极不协调。为此，我们要改善这种局面，把简答题出“活”。克服单一之不足，努力做到在考试中考查考生的思维能力，知识与智能。当然“活”题不是偏题、怪题，不能使考生摸不着头脑、不知所云，“活”的意义在于考试题的内容是实质性的东西，是重点的内容，考生一看即知所问，要正确回答考题，却需要一点真本领、真功夫。

第三节　论述题的编制

一、论述题的特点

论述题是主观性试题的主要代表，在它身上集中地体现了主观性试题的长处。对于综合性考查考生才华有独到的优

势，它能给考生以展示才华的余地。论述题占分量较大，考生解答的时间也较长。它是要求考生作扩展性回答而答案一般比较长的试题。

论述题不是一种单一的试题，而是一类试题，它主要包括以下几种类型。

（1）叙述题。主要是叙述某些事件产生的原因，过程、条件等。如象，试述抗日战争爆发的原因。试述资本循环顺利进行的必要条件，等等。

（2）说明题。主要是说明什么问题的重要性、必要性、迫切性等。如：从事物发展是前进性和曲折性的统一来说明树立共产主义理想的必要性和迫切性，试用社会意识对社会存在的反作用原理说明建设新会主义精神文明的重要性，等等。

（3）评价题。如评左宗棠；写一篇当代英国小说评论，说明它在世界文学中的地位。

（4）分析。如确定糕点在烤后未膨胀起来的原因；了解比喻在散文中的作用，并说明它如何取得特定的艺术效果；分析《红楼梦》中晴雯的性格特点，等等。

（5）批驳。如用实践是检验真理的唯一标准的原理，批驳“共产主义没有经过实践检验”的错误观点。用唯物辩证法的观点，批驳“否认事物内部的矛盾性是事物发展的根本原因”的形而上学的观点。

二、论述题的编制

编制论述题、必须依据论述题自身的特征，扬其长、惩其短。一般情况下，编制论述题时，须注意以下几个问题。

第一、要给考生发挥自己真实水平以较大的余地。

论述题与简答题相比，论述题对于人才的成长来说，考核和反映运用知识去分析和解决问题的能力、在对已学过知识占有的基础上去获取新知识的能力，具有更重要的意义。因此，在编制论述题时，特别要注意发挥它的这一长处；编制适于扩展性回答，更宜于表现考生的独到见解的试题。

第二、要从知识的整体出发，进行综合性考查考生的能力与具体应用基本原理的能力。

简答题虽然也能用来考查考生的知识结构，但它更适于逐点地对知识进行分解式的考查。容易大综合式的考查是论述题的长处。编制论述题，要着重发挥它的这一优势。

人们更多的是使用论述题去考核对原理的运用，其理由是因为容量较小的简答题难以进行这种考核，而这种考核同样是十分重要的。在编制这类论述题时，应把考核的着重点放到利用原理分析、说明、论证的问题上，而不要把核心放在原理本身的叙述解释和对事例本身的介绍说明上，否则，便降低了考核的深度和价值。

第三、要考核考试内容中的重点问题。

考试都应该考核实质性的东西，这是考试的基本要求。但对于论述题来讲，只做到这一步是不够。根据它的特点，绝不能象简述题那样考查一个点，而必须要考查一条线，一个面。很显然，根据试题的抽样，考试试题的内容是由点、线、面的一定组合，它构成了一份试卷的结构。形象点讲，这网状结构什么地方密集，什么地方就是考试的重点内容。论述题是大题。给分量重，有一定深度。因此，它的编制，必须考虑到它的特点，其考核的内容应是考试中的重点问题。

否则，这份试卷将降低效度和信度。

三、论述题的改善

改善论述题主要途径是“客观化”，即评分客观，出题客观予提高试题的效度与信度。

论述题的最大弱点就是评分没有统一的标准，或标准不够客观，试题的得分与评卷员本人的素质，兴趣、爱好等个人因素有很大的关系。许多容量广泛的试题，在评卷员阅卷时很难找到一个公认的标准答案。即使有一个“标准”，其实并不一定就那么标准。它只是用某一位或几位评卷员的“标准”去代替众多评卷员的多样的“标准”。历史上出现的问题在第一节已谈过，是值得深思与觉悟的。由于考生的解答方式是多种多样的，用一种解答法和表述法所写成的答案，与解答法及其表述各异的众多试卷相对照，还是要靠评卷员的主观经验斟酌给分。因此，改善论述题的关键是改善其评分办法，使其评分尽量客观化。

一般讲，一道论述题，考查的是一条“线”或一个“面”。这“线”或“面”通常都可分解为许多要点，这些要点大都以单独拟题考核，将这些编制成小题目，充实考试内容，扩大考试面，增加题量，在这些小题中选留一两道关键性的试题作为考试的重点题目。这样就可以使评分标准比原来要好掌握些，给分客观些。当然，这种分解题目是要有精心设计，认真研究的，否则，就会失去论述题本身的特点。分解千万不能把论述题的长处给分解了。

第四节　作文题的编制

一、作文题的特点

作文是识字写字，用词造句，布局谋篇，运用语言文字表达思想感情的综合训练，是对人的逻辑思维能力、形象思维能力，书面表达能力和思想水平的综合考查。在作文中，往往明显而集中地反映出考生在思想认识，生活经验、知识基础以及语文技巧等方面的情形。考生语文掌握得怎么样，作文可以作为衡量的重要尺度。

作文，一般分为口头作文与书面作文，书面作文又分为命题作文与条件作文。

命题作文，就是主考人给出一个文题，不作任何解释，对写作不附任何说明，让考生独立构思写作。

条件作文，就是给出一篇文章，一则故事，一首小诗，一幅漫画或几句名言作为条件，要求考生循此确定思路，组织文章。或者要求考生写读后感，或者要求按原意改写，扩写、缩写，或者要求由原意生发为文。

根据不同的分类方法，作文有如下几种形式：

1.看图作文。一幅好的图画（照片），是作者对生活反复观察、体验、分析、经过周密思考，集中了最能反映主题的题材而创作的。因此，看图作文，对考查考生观察、体验、分析现实生活，如何选材、组材、确定中心都有很大的作用。

看图作文可以是就画写话，要考生将一幅或几幅内容连

续的通用文字表达出来，也可以是因画抒怀，要考生对画面内容抒发自己的感受；还可以是看画评画，对图画进行分析和评价。看图作文，高低层次的考试都可采用。低层次可以选择内容比较简单的画；高层次则可以选择内容比较深刻的名画。

2. 听写材料。在考试时，让考生听一段或一节，用生动的语言讲述的材料。然后让考生用书面语言复述这些材料、复述有实事笔述和创造性笔述。前者适用于低层次考试，后者适用于高层次的考试。这种作文可以考查考生听写和记忆的能力，深刻地理解中心思想，熟练地运用词语和表现方法的能力。

3. 写读后感。就是让考生阅读一篇文章后，写出感想和收获。这对于考查考生理解文章和写作技巧有很大好处。写读后感没有一定的格式，但一定要认真钻研原作写出真情实感。因此，这种作文有较大的灵活性，能考出考生的阅读能力和表达能力。

例如：仔细阅读下边这篇短文，写一篇读后感

毁树容易种树难

杨树横着可以活，倒着种也可以活。

可是，十个人种杨树，只要有一个毁它，就没有一棵活杨树了。

种树的有十人之多，种的又是很容易活的杨树，却经不住一个人毁它。原因是什么？毁树容易种树难。

（一九八一年高考作文试题）

4. 缩写。缩写是在主考人指定范围内用较原文为少的语言（在字数上应有一个最高限度）表达原文的基本内容。

这种作文能考查考生的综合，概括，提炼的能力。（如1979年高考作文题）。

5.改写。这是要求考生对同一内容用不同形式来表达，这种作文可以考查考生的语言组织能力和想象能力，改写的方式可以多种多样，例如变换人称，变换文体，变换叙述方式等。

6.扩写。与缩写相反，它不是对原文的压缩和概括，而是要求考生对原文的扩展和生发，这种作文有助于开拓考生的思路，考查他们的想象能力和语言表达能力。扩写时可以扩充原有情节，也可以增加新的情节。但必须是对原有情节的合理扩大和补充。扩写的文章一定要严格选择。长文章不宜扩写全文。

7.续写。这是继续补充原文的情节，使原文获得新的发展的一种作文。这种作文可以考查考生的想象力和创造力。续写的文章一般要选用故事性较强的记叙文。

二、作文题的编制。

作文是一种综合的实践活动，它不仅要综合运用语法、修辞、逻辑、写作方法等方面的基础知识，还涉及到立场观点、思想感情、生活经验和知识的深广度等方面。因此，拟作文的文题、看似易事，实则颇需功力；似乎可以信手拈来，实际常常是苦心设计、反复推敲的结果。

编制文题必须注意以下几方面的问题。

第一、文体要根据考试目的和考生现在或将来的需要确定考试的文体要求。

第二、写作意图和选材方向，要根据考生的特点和考试

目的要求确定写作意图和选材方向，选材的方向中有全体考生所熟悉的材料；使考生由文题而出的主意是积极的，有意义的。

第三、考生的发挥余地。要考虑文题的容量，既不怪僻，使人有话可说，又不落俗套，使押题者不能据三、五篇成文而“穿靴戴帽”；既使较低水平的考生能够谋篇作文，又能使高水平的考生有展示才华的余地。

第四、条件作文作为条件的文字。要简短（缩写、改写除外），易读（除兼考阅读能力），含意深刻，有生发的余地；漫画的画面要简单，易于看懂而不生歧义，寓意深刻而有现实意义。

下面选择恢复高考以来的一些优秀作文题，供参考

（1）先天下之忧而忧，后天下之乐而乐（1982年高考作文题）

（2）读林觉民《与妻书》后（一九八二年上海市自学考试作文试题）

（3）千里之堤，溃于蚁穴（一九八三年天津市自学考试作文试题）

（4）有的同学说：“每逢作文，自己常常感到无话可说，只好东拼西凑，说一些空话套话，甚至编造一些材料”。有的老师说：“每次学生作文，我都辛辛苦苦地批改，讲评，但是学生往往只看分数，不注意自己作文中存在的问题，所以提高不快”。请针对上面两段话所反映的情况，联系自己和周围同学的现状，以对中学生作文的看法为中心，写一篇800字左右的议论文，题目自定。（一九八四年高考作文试题）

三、作文题的改善

作文题的改善主要在评分上，要使得评分客观，合理。

由于标准答案和评分规定不可能十分详尽，因而即使它能被众多评卷员所一致的掌握，对同一份试卷也还会得出不同的分数。如上海市招生办曾请一名学生根据一九八一年高考语文副卷中的作文题写一篇文章，邀请13位语文教师按统一的评分标准评分，评定结果，最高给34分，最低25分（满分40分）。

目前对作文的评分，一般采用分解评分法和整体给分法。

所谓分解给分，就是将参考标准答案分解为若干要点，或将正确解答过程分解为若干步骤，每个要点按其与考试目标的关系和重要程度，分别确定其满分量，阅卷时按要点给分。其优点在于能够适当提高给分的客观性，准确性和一致性。其缺点是与教师制定的作为评分依据的标准答案不同的答法，难以按要点逐个对照，评卷时倾向给低分，使另辟蹊径，有独到见解和有高度概括能力，解答简练的考生受到压抑和埋没；而解答文字较多，整体看不得要领，分解看虽废话连篇，语无伦次，却有许多表述要点的词句，这样的试卷反而倾向得高分，从而鼓励考生漫天撒网去罗织重点。

所谓整体评分，就是对考生的答案总体作出评估，判定优劣等级，再斟酌给分数。其优点是：不限制考生的思路和答法，有独到见解和解答透辟的易于得高分；漫无撒网而不得要领的，难以捞取分数。缺点是：给分不易掌握，客观程度低，宽严不一的现象严重，文字的工整、流畅，文辞的华美，常常能够掩盖实际知识的不足；要求阅卷教师知识丰富，

思路开阔，能言百家之言，准确地评分，常需反复阅卷，多次斟酌，如求进度，只能凭浏览的印象给分。

上述各种方法都有自己的优缺点，只有取各自的优点，组合成新的评卷方法，才能使作文题的评分更加客观、科学。

第六章 客观性试题的命题

第一节 客观性试题的概述

一、客观性试题的概念

客观式考试或新法考试（有的又叫“标准化”考试），是按照系统的科学程序组织的，具有统一的标准，并对误差作了严格控制的考试。它可以视为对考试制定出客观而规范性的标准，从命题到实考，阅卷、评分等各个环节都力求减少或避免各种误差，从而测出考生比较真实成绩的过程。只有按照一套严格的科学程序来组织考试，才能有统一的比较标准（即相同的单位和参照点），才能最大限度的减少误差，使考试尽可能准确可靠。

客观考试只有七、八十年的历史。1908年，法国心理学家比纳和西蒙出版了《比纳—西蒙智力标准》，为智力测验开辟了新时代。1904年，美国心理学家桑代克出版了《智力与社会测量原理》是教育测验方面最早的教科书。其后出版了《儿童书写标准》，使测验方法向客观化大大迈进了一步。1908年和1909年美国教育家约·斯通和斯·柯斯发表了客观化的算术测验，把教育测量原理成功地应用于具体学科的考试，后来产生的斯坦福成绩测验是第一个主要的客观式成绩测验。麦柯尔（A · Mcall）于1920年发表文章，提倡仿

照标准测验的命题形式编制试卷。于是欧美国家的一般学校广泛采用了这种考试方法。共和国成立以前，我国一些教育家也提倡用新方法考试来测量学生的能力，改进教学方法。1931年成立了“中国测验学会”，并发行“测验杂志”，师范院校开设“教育测验与统计”课，编制了各种心理和教育标准化测验。共和国成立后，由于学苏联，照搬“五级成绩考评方法”，全盘否认了新法考试，使我国对新法考试的研究中断了三十多年。近几年才开始这方面的研究。

客观性试题和主观性试题一样，它不只是一种试题类型，而是具有某种共同特征的试题类型的总称。单看它的正确解答是否唯一，判卷给分是否客观，那么，正确解答唯一，不论由谁判卷都只能给出同一个分数的，叫做客观性试题。从试题本身的特点来说，客观性试题大都是“供给式”或“固定应答式”的，即问题本身就给出了一种或几种固定的答案，考生的解答就是对已给答案正确性的判断。对正确答案的选择、判断，选择对了就给满分，错了就给零分。

客观性试题种类比较多，大概有如下几种：

①填充题，如：

产业资本连续循环的公式是________。

一切权力属于人民的代表制民主有两条根本原则：

（1）________（2）________

民族形成的一般过程和规律是：由____→____→____→民族。

②改错题，如：

今年的生产任务提前实现了。（完成）

Good manners should be of served whether one eats in a restanrant or in home。

(A: manners; B: whether; C: one; D: in)

(D) at

③判断题，如：

党的优良作风即理论和实践相结合的作风，和人民群众紧密联系在一起的作风，批评和自我批评的作风。（对吗？）

批发价格和零售价格的差额就是商业纯利润。

④选择题，如：

度是__

1. 质和量的统一。
2. 事物的质所规定的量的活动范围。
3. 事物的量的极限。
4. 保持事物质存在的量的限度。

马克思主义哲学是________

1. 研究世界观的理论体系。
2. 研究世界本质及其发展的一般规律的科学。
3. 无产阶级认识世界和改造世界的思想武器。
4. 完备的、严密的唯物主义理论体系。
5. 自然、社会和思维知识的概括和总结。

I am sune that _____ you said is wrong

A with B all C this D what

⑤分类题，如：

给出的作品（A－E）按I到V所列的特征分类。

A 屈原 B 子夜 C 离骚 D 巴黎圣母院 E 安娜·卡

列尼娜

Ⅰ 现实主义　B、E

Ⅱ 浪漫主义　A、C、D

Ⅲ 诗歌　C

Ⅳ 小说　B、D、E

Ⅴ 戏剧　A

⑥配对题

1) amusement (b)　a) ability to do some thing well

2) healthy (c)　b) enjoyment

3) virtue (d)　c) strong and well, not often ill

4) skill (g)　d) goodness of character

5) valuable (f)　e) news or knowledge given

6) information (e)　f) useful, of great value or use

7) experience (a)　g) knowledge or skill coming from practice

1) 低栏 (C)　A caress

2) 抚爱 (A)　B native land

3) 雇农 (E)　C low

4) 故土 (B)　D clap one's hands

5) 鼓掌 (D)　E farmhand

6) 混乱 (G)　F soul

7) 灵魂 (F)　G coupusion

二、客观性试题的优缺点

客观性试题的主要优点：被试者解答题时，方便简单，因题量较大，考查面广，有利于全面了解应试者掌握知识的情况。一般客观考试试卷题目多达50－100个，甚至更多。因为它不象主观性试题那样要求被试者自己组织答案，并用文字表达出来，不象主观性试题那样把大量时间用在书写上，而是用大量时间去思考、去阅读。由于试题多，会使随机误差的影响相互抵消。试题取样范围广，可以促使被试者按照考试大纲和教科书规定的内容全面复习。可以在一定程度上削弱搞题海战、猜题、划重点复习范围等对考试的影响。

客观性考试答案明确，因而评分比较客观，客观考试题的种类按记忆反应的不同可分为再认式试题和再生式试题两大类。再认式试题包括被试者把学过的知识重新出现在被试者面前，让被试者辨认或加以排列、组合。是非题、选择题、配合题、顺序题等都是再认式试题。从而，客观性考试的评分不受评分者主观因素的影响。任何评分者评定同一份试卷都会得到相同的分数，同一评分者在不同的时间评定同一份试卷也会得到相同分数。被试者的分数几乎由考生本身解答试题的正确程度所决定，而不是由评分者所控制，评分者的兴趣、情绪、知识程度不影响考生的分数。提高评卷的效度与信度。

客观性试题经过预考，使考查比较准确有效。客观性试题施考前要经过预考、分析和修正筛选。对被试者的水平有比较正确的要求，为此，难易度比较适合，区分度较高。

客观性试题计分标准化，使得被试者的成绩便于区分和比较，有利于择优，有利于实现考试阶段的现代化，使用电

子计算机进行评分、计分和统计分析。可以大量节省人力、物力和时间。

但是，客观性试题发展到今天，也不是完美无缺的，还有不少弊端。主要出现在以下方面：

客观性试题答题比较简单，被试者的表达能力、组织能力、独创能力、创新能力、发散思维能力、逻辑推理能力等很难反映出来；也难对被试者的综合能力和统合能力进行考核。另外，虽然一般的猜题打题现象可以避免，但被试者通过对一定考试方法的熟悉和研究而凭借经验投机取胜的可能性还在一定程度上存在。客观性试题的编制比较复杂，比较困难，需要较高的命题技巧和较长的命题时间。

三、主观性试题与客观性试题的比较

主观性试题与客观性试题，是两类最主要的题型。全面比较它们之间的各方面的因素是非常必要的。

题型 / 因素	主观性试题	客观性试题
考查方向	一般是从具体知识的整体上进行综合式的考查	具体知识分解为许多较为单纯的要点、要素，进行分解式的考查
试卷的解答	要求考生写答案，容许考生写出草率的和无关宗旨的答案	要求考生选择正确的答案，容许考生猜测作答。
试题量	一般有几道、十几道。	一般有几十道、上百道。
思考的方式	扩展性思考，是构思文章。	集中式思考，是判断已给答案的对错。

续表

题型 因素	主观性试题	客观性试题
取巧的因素	以文辞的优美掩盖知识的不足。	凭机遇猜答案。
知识考查	侧重于考查知识的深度，样本需从答案中确定。	侧重于考查知识的精细和广度、样本由拟题者决定。
时间的消耗	花时间于思考和书写。	花时间于阅读和辨识。
主考者的最大困难	试卷的评阅。	试卷的编制。
绩分分布	评卷者控制。	由所答对错决定。
考试结果	反映考生解答过程，正确的程序和错误所在。	考查考生的最后结果。

两类考试各有长短，不能说哪类绝对优于另一类，只能说在某个方面，从某个角度来看，哪类要优于另一类。两类试题，都有很强的适应性，不能说哪类只能适于某些学科，某些知识的考查，对其他学科，其他知识的考查无能为力，只能说在某些条件下，为了某种目的的考试，使用哪类试题更合适。

第二节　是非判断题的编制

一、是非判断题的特点

是非判断题在客观性试题中，是应用得比较多的一种试题，它仅次于选择题。它给出一个含义完整的命题，让考生判断这个命题的是非对错——即是非判断题提供正确和错误两个答案给考生选择。正确答案可写成“正”、“真”、“是”、“对”等不同形式，错误答案亦可写成“负”、“伪”、“非”、“错”等相对的字。

在某种程度上，是非判断题就是选择题。只是形式和方式不同于选择题而已。为此，要把它与选择题严格分开，也是较困难的。实际上，它们有许多共同的东西，但它们毕竟不是同一事物，存在差异。因而，有研究它的必要。

一般认为，多项选择题是多个是非判断题的汇集。而它的陈述因素要求统一的题干相搭配，各项之间在表述上要相协调，往往不如各个独立的是非判断题的陈述方便、准确。同时，任何多项选择题都可改变成为单项选择，因而不主张使用多项选择题代替是非判断题。

是非判断题，常只有一句话，一般讲凡是比较简单的陈述句，其编制比较容易。试卷取择宽广，评分客观，容易，快捷。

是非判断题的选择答案只有两个，考生经猜测而得分的可能性高达50%；容易推测正确答案。可靠性相对减弱。若设计不严谨和使用不恰当，便会使考生过于集中学习零碎的

事实和着重低层次的认识。

是非判断题考查的角度是各种各样的，其性质也同样如此，以下是不同性质是非判断题的例子：

题　　根	答案	性质
1、商品经济的国家里，大多数公民不重视个人利益	正或负	推论
2、因为：0.5>0.25，所以$\log_{0.5}0.5>\log_{0.5}0.25$ ⇒ 1>2	对或错	因果
3、五个测验分数（6、7、2、4、5）之中偶数是2	真或假	混算
4、以摄氏温度衡量热能，40℃水的热单位是20℃热单位两倍	是或非	比较
5、大礼堂的倒塌，是万有引力的例证	是或否	证据
6、当某人掌握了修辞和语法，他更能写出好文章	对或错	条件

二、是非判断题的编制

是非判断题的设计必须注意它的特性和设计技巧，必须遵守以下要点：

第一，设计题必须能够从每一题辨认出代表重要的概念的中心意思。如果一个句子含有一个以上的观念，读者就得对它们都加以注意，结果便可能读错或误解这个句子。

第二，每一试题必须是肯定正确或肯定错误，不可以模棱两可。但必须有一定的连贯性。否则，判断二字就体现不出了。

第三，语言要简单，应用双重否定和复杂易费解的句子可能导致混乱，意义含糊，对考生的干扰严重。语言的简化

有助于降低意义的含糊，语言如果变成了重要因素，那末所考查的就可能是阅读的能力而非学科的知识了。

第四，每一试题不要在字面里提供答案的线索或指导解答。正负答案的次序，不应有规律或模式，以免为被试者猜对。试卷内的正题和负题都应该有数目相近的字数和结构相当的句子。

第五，每一试题必须是考核重要概念，而不是考核零碎知识，一般知识或普通常识。更不能直接从教科书里抄袭句子。如果直接从教科书抄袭正确的句子，或只变动一个字使句子发生错误，那么这个考试问题所考的也许大都是机械记忆而非是理解的了。

第三节　选择题的编制

一、选择题的特点

选择题是提出一个问题或写出一句不完整的话，接着给出这个问题的几个答案或这句话的几种补充说法，给出的答案或说法中，有的是正确的，有的是错误的，让考生把其中正确的选择出来。有时，也让被考者选择一个唯一不正确的答案，或只选一个最好的答案，一份精心设计的多项选择题的考卷，它能有效地考核应用能力，分析能力，综合能力。选择题的被选答案有不同的难易程度和反映被试者不同的错误观念，它有“诊断”价值，避免考生答题时“吹嘘”和“掩饰”，同时也避免了对书写能力低的被试者的惩罚。选择题最突出的优点在于评分简单，客观、准确，可以用计算

机进行，它所组成的试卷，题量大，覆盖面广。

选择题是客观性试题中应用最广的一种题型，深受重视，不少国家和地区把客观性试题与选择题等同起来。

一般而言，选择题是由两个部分构成的，提出的问题或不完整的句子，通常称为“题干”，给出几个答案或补充说法，通常称作“题枝”或“备选答案”。备选答案中又分为“正确答案”（最优答案）和错误答案（分心答案、迷惑性答案）。

例如：

1、The business is risky. But____ we would be rich.

A、should we succeed

B、would we succeed

C、might we succeed

D、could we succeed

2、时间的特性是________。

1）多维性　2）三维性　3）一维性　4）客观实在性

3、How much work is required to compress this gas adiabatically to a volume of 16.8 liters?

A）330 joules

B）680 joules

C）7.81×10^{3}joules

D）223 joules

E）91 joules

例1、2都是一句不完全的句子，给出四种补充说法，其

中例1中只有A是正确的，例2中只有3）是正确的。而例3则是一道计算题，它给出了5个答案供选择。

选择题按备选答案中正确答案的个数，可分为单项选择和多项选择题。单项选择题，在已给出的备选答案中，至少有一个是正确答案，如上面的例子。

二、选择题的编制

选择题是客观性试题编制中最难的一种，尤其是要考核认识能力的选择题，更需要讲究设计技巧。构思方法也是十分重要的，它是编制选择题的关键。下面介绍构思方法。

直接法　根据大纲和教材所要求掌握的基本概念，基本规律和基本技能进行构思。

如要考查$y=\sin x+1/2\sin 2x+1/3\sin 3x$的周期

可设计为：$y=\sin x+1/2\sin 2x+1/3\sin 3x$的周期是____

（A）2π　　（B）4π

（C）6π　　（D）$2\pi/3$

（正确答案为（C））

变形法　若一个问题可以从不同的侧面，不同的角度来叙述某一事物观念或事物概念、事物规律的，可把这些正确答案支变形成迷惑支，只保留一个正确答案。

例如关于力的合成和分解，下列说法不正确的有：（1）不可能将一个力沿它的垂直方向分解出两个力；（2）合力一定大于分力；（3）合力的数值随着两力间夹角的增大而增大；（4）共点力的合力一定为0。（答案：（2）（3）（4））

若将上题中“下列说法中不正确的有”改为“下列说法

中正确的有”就成了我们所要设计的题目，答案只有（1）了。

改选法　常规的填空题，计算题、证明题、作图题、说理题等都可改选成选择题。其方法是把原题中的正确答案改造选成几个迷惑支。

如求1／2与1／3之和？就可以设如下几个答案：（1）1／6　（2）2／3　（3）2／5　（4）5／6　（正确答案（4））

罗织法　在平时的解题过程中，学生因不同的原因而造成错误，如审题不细；思路不清，根据不足，以及运算不对等，搜集这些典型错误并进行罗织就可编成选择题，这有利于比较辨析，纠正学生的错误。

如：decline（v.谢绝），refuse（v.拒绝　不肯）

decline表示“婉言谢绝”，语气比较委婉，它不能用人作宾语。

refuse表示“拒绝提供某物或者拒绝按别人的要求去做某事”。因此就可以编拟如下试题。

例如：He was afraid he would have to ____ her invitation to the party

A、refute　B、refuse　C、return

D、ignoce　E、decline　（E对）

合成法　选择支用迷惑支加“以上答案却不对”而合成的选择题也具有一定的巧妙性。

选择题由于其自身的特点，在编制时还必须特别注意以下几点：

第一，项目的内容要尽量包含在语干之中，语干要明确表达一个问题。必须简短扼要，避免不必要的重复和混乱，

没有无关的材料。

第二，要使一切可能的反应在语法上都是正确的。应当把语干和备选答案编制得使任何一个备选的答案读起来都带有语干。而且语法正确。如果这些备选的答案在语法上与语干不协调，它们对较优秀的学生说来就不是真正备选的答案。

第三，所有不正确的备选答案都必须似乎是合理的。假使备选答案中有某一个似乎是不合理的，结果就增加了碰巧答对的机率。被试者对于简单的回忆项目所作出的那些不正确答案，当这些项目以后被用作多重选择的项目时就成为一些备用的不正确的反应了。

如何提高选择题的可信程度，这是客观性考试的首先要努力解决的问题，然而，从客观性考试的现状来看对这一问题解决的还不很理想，有待于研讨。

三、多项选择题的编制

多项选择题，一般给出几个备选答案，其中可能有两个直到几个备选答案是正确的答案，但并不告诉考生每一试题正确答案的个数。

编制多项选择题时，除了正确的答案个数不定外，必须注意以下几点：

第一，迷惑性，迷惑性是选择题的应有特点，答案似是而非，似非而是。因而编拟的每个选择题支都应该有被选择的可能，而且需经认真分析题意；深入理解，完成推理，演算等之后，才能排除迷惑支。这种迷惑性的多项选择才有价值。

第二，思考性，选择题最大缺陷是不便于弄清受试者分

析题意，解决问题的思路，但在编拟选择支时却应注意题目思考的价值，应使具有错误思维方法的受试者在得分上占不到便宜。

第三，似真性，多项选择的每个迷惑支，要求能反映受试者知识中存在的某种缺陷，不应是一句随意写上的结论（或随意填上的数字），决不能为了求方便，编拟了一两个正确选择支后，迷惑支就随意设计来凑数。这是极不严肃也不科学的态度。

第四，对比性，多选择题的多选择支是对同一问题进行多侧面的描述和多角度的分析，前后应有一定的联系，便于对比分析和选择。毫无关联的现象的事例达不到这个目的。

多项选择题与单项选择是互通的。单项可改为多项，多项则可改为单项，就多项改为单项的办法一般有以下几种。

1、分解为多题。把某些答案放在题干中或舍弃掉，来考其中的一两个答案。这样一道多项选择题就可以分解成多个单项选择题。

如：哲学的基本问题是：

（1）唯物主义和唯心主义的关系问题。

（2）意识和物质的关系问题。

（3）辨证法和形而上学的关系问题。

（4）思维和存在的问题。

（5）事物之间的矛盾关系问题。

将（2）（4）分别与（1）（3）（5）组合，就得到多个单项选择题。

2、采用“以上都是”作为备选答案。

例：若f（x）为（$-\infty$，$+\infty$）上的任意函数，则

$F(x)=f(x)-f(-x)$ 是

(A) 偶函数　　　　(B) 奇函数

(C) 都不是（加的）　(D) $F(x)=0$

3、将多项选择题的各项应答支放到题干中，把各项的几种组合作为备选答案。

如：$f(x)$、$g(x)$、$h(x)$ 均为奇函数，则（　　）中给定的函数是偶函数。

(1) A、B、C

(2) C、D、G

(3) B、F、G

(4) A、F、G

(5) E、G、B

A　$f(x)g(x)h(x)$　B　$(f(x)+g(x))h(x)$

C　$f(x)+g(x)$　D　$f(x)+g(x)+h(x)$

E　$f(x)-g(x)$　F　$(f(x)-h(x))g(x)$

G　$|f(x)-g(x)h(x)f(x)|$

4、改变问题的问法，让考生选择错误的答案。

如 $\frac{252}{420}=?$

(1) 60%　(2) 0.60　(3) 21／35

(4) 29／36　(5) $\sqrt{1-(\frac{4}{5})^2}$　答案为 (4)

四、选择题的修改

选择题试卷编制好后，经过预试，分析试卷和每一试题的质量，不合要求的，进行适当的修改必要的调整。

选择题试卷预试后的项目分析，主要是信度、效度、难

度与区分度。但是选择题这种题型有其自身的特点，对它的上述指标的分析除与其它题共同之处外，还有一个特别重要的分析，就是误道分析，在不同程度上讲，误道的研究，误道设计质量的高低，决定选择题的一切。对误道的分析办法，常是将高分组（成绩最高的27%的考生）和低分组（成绩最低的27%的考生）选择各误道答案的情况统计出来。记录下来，借此比较各误道答案的优劣。下面举例说明。

试题分析式样

f（难度）：0.34　　D（区分度）：0.45

题号：15　矛盾斗争的绝对性，矛盾同一的相对性是指：

*（1）前者是无条件性，后者是有条件性。

（2）前者是主要的，后者是次要的。　（2－10）

（3）前者是积极的，后者是次要的。　（2－13）

（4）前者是有用的，后者是无用的。　（1－5）

放弃不选　（0－3）

其中：①误道答案（2）后面括号内的数字（2－10），表示高分组中有2人选（2），低分组中有10人选（2）。其余误道答案后面的数字类似。

②放弃不选后的（0－3）表示高分组没有人对该题放弃不答，低分组中有3人对该题弃权，没有作答。

修改的程序是：首先，将试题按区分度大小分组；D＞0.30为通过组，0.30＞D＞0.10为修改组，D＜0.10为淘汰组，其次，对修改组的试题，逐个进行修改，再次，分析淘汰组质量低的原因。如果是整个试题设计得不好，则予以淘汰，并再编制顶替试题。试题修改，试卷调整后，还应组织预试，最低也要对修改过和重新编制的试题进行预试。直到

符合要求为止。

第四节　其它客观性试题编制

一、填充题

填充题分为填图题和填空题两种。

填图题，就是给出一个不完整的图形（如南方古道路线图），要求学生填写出不完整部分（如注出路标、地名等）。通常用于历史及地理考试中。

填空题，就是给出一个不完整的句子，要求考生把不完整的部分补充上去。

填充题分为选择填空题和自由填空题。

如，斯大林1913年给民族下的定义所指对象________

1、古代民族　2、资本主义民族　3、社会主义民族　4、所有民族

He gathered a bouquet ______ flowers ______ his mother from the flowers ______ the garden while he was waiting.

Her friends were kind ____ her ____ her time ____ trouble.

马克思主义理论产生是用以教育无产阶级，这样就使无产阶级理解了________理解了________理解了________这时他们就变成了一个________。

前者是选择填空题，后三题是自由填空题。可以看见，选择填空题，很相似于选择题。实质上，两种填空题均可改

为选择题。

编写填充题，要注意以下几点：

1、自由填空题提问要求尽可能明确，使空白中应填的问题具有单一性。

2、空白中要填写的应是给出的这句话的关键词语。

3、如果一句话中留下多处空白，给出的文字对要填写的文字应有足够的暗示，或者说，留出的空白不应该使原文面目皆非，使考生也难以想出原文的真实面目。

4、同一试题的多处空白，考查同类问题不同试题的空白，主动在卷纸上留出的空白大小应当一致，不要给考生选择填写问题的不应有的暗示。

二、分析判断题

分析判断题是给出一个断言和说明这个断言的一条理由，让考生分析判断这个断言的正误、理由的正误，理由能否说明断言，并将判断的结果用规定的符号表示出来。要比是非判断题更便于应试者理解和运用，考查的准确性更高，猜测得分的可能性较小，但考生答题费时较多，试题取样没有是非判断题广泛。

如规定：认为断言和理由都正确，并且理由能够正确说明断言，记作“A”，认为断言和理由本身都正确，但理由不能正确地说明断言，记作“B”，认为断言正确，理由错误，记为“C”，认为断言错误，理由本身正确，记作“D”，认为断言和理由都错误，记作“Z”。

例1，如果一个数能被4整除，那么它也能被2整除，20能被4整除，所以，20能被2整除。

该题的理由和断言都正确，又能正确说明断言，因此答案应该写“A”。

例2，所有的事物是运动着的，因此有些战争是非正义战争。题中的论断是对的，理由本身也对的，但没有正确说明断言，答案是“B”。

例3，大发明家瓦特是没有受过高等教育的，因此科学家有没受过高等教育。该题的论断是正确，但理由是错误的，因此答案应该写“C”。

例4，所有的哲学系学生必修逻辑课，哲学系学生是文科学生，因此，文科学生必修逻辑课，该题论断是错误的，但所列理由却是正确的，答案是“D”。

例5，工人就是矿工，因为没有耕耘，就有收获。题中的论断和理由均错误，因而答案是“Z”。

编制分析判断题时，除是非判断题的编制要求外，还特别要注意，理由与论断的搭配。要避免理由明显正确或明显荒谬，否则，试题只须判断论断或是非了，变为是非判断题，要避免理由与论断说明是毫无关联的两回事，否则，试题只须分别判断论断和理由了，同样变成了是非判断题了。

三、改错题、分类题、配对题

改错题：就是给出一个错误的句子，让考生找到错误处，并把错误改正过来。改错题比填空题“活”，它侧重考核对知识的理解和运用。但它的编制受到较大的限制，许多知识难以编成改错题来考试。但改错题对考核语言知识有独特之处，如：

1. 只有努力工作，才能取得更好的成绩。（“只有’

改为“只要”）

2. 四个现代化的宏伟蓝图一定能实行。（“实行”改为“实现”。）

3. 烈士们的革命英雄主义行为是何等崇高啊。（“行为”改为“精神”）。

分类题，就是给出几个事物（或特征），列出相关的一系列特征（或事物），要求学生按所属关系进行分类。如：

1. 请将给出的书（A—G）按①到④所列特征分类。

A《电动力学》　B《高等代数》　C《热学》

D《生物化学》　E《光学》　F《有机化学》

G《数学物理方法》

①数学类　B、G

②物理类　A、C、G

③化学类　F、D

④生物类　D

2. 请将给出的单词（A—F）按①到③所列的特征分类。

A. Januany　B. jest　C. musical

D. teacher　E. reedy　F. make

①名词（n.）　A、B、D

①动词（v.）　B、F

②形容词（adj.）　C、E

编制分类题，特别要注意相容的分类方法的考核，与不相容的分类方法的考核，给出的被分事物要相近，才能考查其深度。

配对题提供多个题意和多个答案。被试者需要把每一个题意配上他认为正确的答案。配对题最大的长处是节省试题

的空间，把多个问题或题意与答案串起来。以配对形式考核各种认识能力。尤其是对名称、日期、地点、事实、事件、名词、定理、规律、类别、符号、标记等的联想辨认能力最为适合，配对题最大的局限性是编制一连串性质相近而考核所要求的教学目标的题意和答案。这就需要编制试题者掌握设计试题的技巧了。若不设计严谨，则配对题又会成为只考核事实的记忆和联想，而忽略高层次的认识能力的考核。

配对题的例句：

把右列的字母填入左列的括号内：

( ) 1. 十八世纪初　a. 发现美洲
( ) 2. 十八世纪末　b. 美国首位太空人
( ) 3. 格兰　c. 美国之父
( ) 4. 哥伦布　d. 首届美洲大陆国会
( ) 5. 华盛顿　e. 美国宪法章程草拟成功
f. 发明电话

( ) 1. present　a. stop
( ) 2. prese　b. one of a pals
( ) 3. mate　c. book
( ) 4. puit　d. give
( ) 5. volume　e. hug
( ) 6. halt　f. result

优秀的配对题是较难编制的。除非编得很好，否则就有鼓励机械记忆的倾向。它的编制应符合它的特性。讲究设计技巧。下面几点建议可供参考。

第一，每一个配对题以仅含有相同的材料为限。否则就可能增加用排除法求得某些答案的机率。

第二，每组配对题的各个答案需要对各个题意有大致相同的误道作用，每一项目只许可有一个正确的配对。

第三，对应的备选答案必须按某种逻辑的次序排列起来。依照字母或其它次序排列可以减少被试者的抄写工作和完成练习所需要的时间。

第四，配对的各组不要太长。长的配对练习耗时过多，可能造成混乱。十对至十五对就可能够用了。

第五，每组配对题的题意相同和答案数目不应该相等。这样就可以避免“排除其它的方法”求得某些答案。

填空题、改错题、分类和配对题，没有选择题和判断题那样广泛地应用，人们也很少用这几类试题来编制一份完整的试卷。但是，这些试题有其自身的长处，在特别的环境下，它们均有用武之地。

第五节　客观性试题试卷的编制

一、客观性考试的种类

客观性考试种类繁多，名称各异，从不同角度可做出不同分类。

按考试性质分：有成就考试与能力倾向考试。成就指的是经过一定的教育或训练后所学到的东西，是在一个比较明确的，相对限定的范围内的学习结果。成就考试又可分为单科成就考试与综合成就考试两种。而能力倾向则是指学习的能力，是在给予适当的机会时，获得某种知识或技能的能力，并且，此种能力是在一定的遗传素质基础上，生活中各种经

验积累的结果。

按考试要求分：有难度考试与速度考试，难度考试包含多种不同难度的题目，由易到难排列，其中，有些题目几乎所有考生都解答不了，但作答时间较为充裕，因此，它考查的是解答难题的最高能力。速度考试则是题目虽然较为容易，但数量多并且严格限制时间，它主要考查反应速度，而多数考试是以上二者的综合。

按考试材料分：文字考试与非文字考试，前者所用的是文字材料，考生用文字作答，后者所用的材料是图形，实物等，无需使用文字作答。

按考试对象分：有个别考试与团体考试，前者只是考一个人，而后者可同时考评多人。

按考试时机分：有进展性考试与总结性考试，前者在教学进行中实施，后者在教学结束后实施。

按解释分数的方法分：有参照常模的考试与参照标准的考试。前者，是把考试分数与常模作比较，后者则是把考试分数与某种标准作比较。

按考试功能分：常见的有选拔考试（如高考）；安置考试（如按能力和知识水平分配工种）；准备性考试（主要用考查考生对于完成某项学习或工作任务是否作好了准备，即是否具有所要求的最低能力，是用作预测水平的考试）；诊断性考试（主要用来确定学生学习困难之所在）；证书考试（是总结性的水平考试，主要用于对正规或非正规的学历给予承认，或用于为某种职业者发执照）；用作研究工具的考试（国际教育成就评价大会举办的一系列标准化考试）。

客观式考试种类繁多，名称不同，但它们有相同的要素，

有考试大纲或考试指导书，用以规定考试的目的，内容、要求、题型、方式及计分方法，依据考试大纲制定出一个"命题双向细目表"；试题要经过预测或调试，以数量化要求来评估试卷和试题的质量；考试的组织实施过程要统一而规范，综合各科考试成绩时要用标准分数；要提供为解释和评价考试分数用的常模（指考试集体的平均分数；标准差等相对的地位量数）和考试误差。

二、客观性试题试卷的编制

试卷的编制主要从两方面入手，一方面要将教学大纲在知识和能力两个方面具体化。从而制定教学目标并使之可以度量，另一方面要加强编制题目的科学性。

客观性考试的试卷编制主要遵循以下六个环节：

首先，明确考试目的，考什么是考知识还是考能力？哪方面知识？哪方面能力？考谁？被试者的年龄、智力水平，知识结构及文化背景等。为什么考？即考试的功能是什么，是选拔性还是教学分班性的，是评价教学质量还是诊断学习困难程度等。功能的不同使之题目的内容，形式以及难度，区分度的要求发生差异。

其次，制订考试大纲，包括：哪种类型的考试，考试的性质是什么，适用于哪些团体，考查哪些知识与能力，试卷包括几部分，各部分的比例如何？相应的重点是什么，采用何种题型，共有多少题目，用何种方法计分，与其他考试类型的关系，等等。大纲是命题的准则和依据，也是被试者应考的参考。

再次，拟定编题计划，这是具体编写试题的重要环节，

而作一张二维细目表（也称双向细目表），定出试卷所要考查的知识和能力，以及每一种知识，能力的相对重视程度。

知识即是某一学科的各个课题，能力指的是通过教学在认识行为上要达到的目标。美国心理学家及课程权威B·S·布卢姆等人把学习的认识活动分为识记，理解、应用、分析、综合、译价等六个层次。

“识记”在这里指特殊事例和一般事例的回忆，对方法和过程的回忆。鉴于考查的目的，这种回忆仅是将原来的材料再现。尽管允许再现的材料有一些改动，但相对说来必是无关紧要的部分。识记目标主要测重于记忆的心理过程。即知道事物的名称，具体事实。处理具体事情的方法和程序以及有关方面的基本概念，原理和法则。

“理解”在这里指的是低层次的理解。指一个人知道所交流的事物是什么，并能初步应用该事或该物的观念或材料，这种应用是不必和其它材料联系起来，也不必充分了解它们的内涵的，它包括解释事实和原理、法则、解释语言材料或图表，图象；对要点作出分类、摘要、归纳；将材料从一种形式转换成另一种形式。根据材料的内容推断未来的结果，确定方法和程序等。

“应用”指将抽象概念应用于特殊或一般的环境。这些抽象概念可以是一般的观念，程序的规则或一般化的方法，也可以是必须记住并加以应用的原则、观念、理论。它包括将概念、原理、法则、定律等应用到实际中；解答试题，绘制图和图象；方法和程序的正确使用与演示等。

“分析”指将某交流材料分解为组成的因素或部分，这样其观念的相应层次就清楚了，或所阐述的观念之间的关系

也就明确了。这些分析旨在阐明交流材料，指出交流材料是如何组成的，表明交流材料是怎样起作用的，其基础和安排是怎样的。它包括对各组成部分的辨认；对各部分之间相互关系的分析；对把各部分组合起来的原理法则的识别等。

“综合”指将各因素和成份合在一起，从而形成一个新的观念体系。这包括综合评价部分，组成元素等的过程，并包括将这些组成以前所不甚清楚的一种模式或结构。包括综合运用知识以解答问题；写出组织得很好的作文或其他有创造性的作品；制定计划或提出方案等。

“评价”指根据一定的目的，对材料和方法的价值作出判析。从质和量上对材料和方法是否达到准则和要求的程度作出判断。这是认知学习的最高水平，包括评价书面材料中的逻辑一致性。评价证明结论的材料是否适当、充分，评价作品的价值等。

由于B·S.布鲁姆的认知分类已为多数人所接受，所以，一般都依据上述认知性行为目标编拟学科试题计划——二维细目表。

如下表，即为一个初二平面几何第一学期期末考试的考试蓝图。

在制定二维细目表前，要对教学大纲和教材进行深入细致的研究分析，以确保分数合理，比例恰当，当然，根据教材内容不同，可以采用不同的认识层次，并不是每个学科，每项内容都要按这六个认知层次命题。

二维细目表是科学编制试卷的设计图，其主要用途在于指出应该编哪些种类的题目和各编多少，并可按表中百分比确定每种题目应占多少分数，同时，二维细目表不仅是编制

知识内容	学习水平						Σ
	识记	理解	应用	分析	综合	评价	
直线、射线、线段	1	1	1	1			10%4
角	1	1	1	1			10%4
相交线、垂线		1	1				5%2
平行线	1	1	2	1	1		15%6
命题定理证明		1	1				5%2
三角形	1	1	1	1			10%4
全等三角形	1	2	3	1	2	1	25%10
等腰三角形	1	2	2	1	1	1	20%8
Σ	15% 6	25%10	30%12	15% 6	10% 4	5% 2	100%(40)

客观性考试试卷的依据，也可以对没有拟定编题计划而编制的试卷进行检验，看其是否符合要求。

第四，编写审订题目，编题计划（二维细目表）确定后，即可据以编拟试题。一般考试机构的编题，都是组织内部专家命题和对外征集题目相结合。所有的题目都要抄写在命题卡上，不但要写清试题的正文及答案，还要说明所考的知识，能力范围，当试题积累到一定程度时，即由审题委员会按一定质量要求，审查每道试题在学科内容上的准确性，所考试能力范围及命题的技术性，选出初审合格的试题。

第五，对试题进行预试，试题虽然初选出来，但还不能凭主观来判断其优劣，必须经过实际预试，以获得客观性资料，为进一步修正和筛选提供依据。预试要在具有代表性的

预试对象中进行。并对其结果进行统计分析。以确定题目的难度、区分度，对选择题还要确定几个备选答案的适宜度。

第六，拼配试卷。经过预试和题目分析，对各个题目的性能已有可靠的资料作为评价的根据，这就可以选出符合要求的题目加以适当编排，组合成试卷。

为增加实际效用，在多数情况下，试卷需要副本，以便于使用，试卷的各份副本必须等价。其编造手续是，将所有适用的题目按难度排列，其次序1、2、3……，如果要分成n个等价的试卷。就应按下列分配题目：

A_1	1	2n	2n+1	4n	4n+1	……
A_2	2	2n-1	2n+2	4n-1	4n+2	……
⋮	⋮	⋮	⋮	⋮	⋮	⋮
An	n	n+1	3n	3n+1	5n	……

例如：n=3　A_1　A_2　A_3（为副本）

A_1	1	6	7	12	……
A_2	2	5	8	11	……
A_3	3	4	9	10	……

这样做可以使副本之间在难度上基本平衡，从而获得大体相同的分数分布。

另外，客观性考试试卷的编制还需要经过试卷使用的客观化（规定相同的施考条件，统一试卷指导用语和时限等；制定正确答案和评分规则，制订解释分数所需的模式标准），收集效度和信度资料（对试卷进行质量鉴定）以及编写考试手册和指南（分别供考试主持者和考生用的有关试卷的说明书）。

三、客观性考试的改善

客观性考试的缺点有时使得我们在本应采用它们的时候却不去用它。编制大量考查项目的任务对最有经验的主考人来讲也是繁重的，往往觉得是举行考试的时间了，却没有预先把考试准备好。于是他只得应用非客观性的考题，而不管他们是否有效。要编制合适的客观考试项目和客观考试的问题是很重要的。为此，必须对客观性考试进行一些改善。下面提出几点建议性的意见。

首先，考试要采用所考学科的重要成果。施考者在编制考试项目时，必须确信学科的显著特点已受到重视，考试必须考核被试者已学过的内容。必须保证考试能反映出学科重点的适当比例。同某一些项目有关的考试项目究竟应占多少，需参照这个学科对于这个课题的重要程度而定。

其次，考试题的编制必须考虑它所服务的目的以及它在施行时的条件。

第三，一个特殊类型的所有项目都必须放在一起，在考试内采用多于一种的项目是可以允许的，也是十分需要的，有助于破除考试的单调，但是这些项目混杂在一起时，就可能导致纷乱，被试者就要付出额外时间，因而减弱了客观性考试所容许的广泛取样。

第四，考试项目的阅读难度必须降低，除非考试的目的在于考试阅读的能力和词汇，否则阅读难度要降低到受测验者都能了解题的意义。考试的项目必考查学生的学科知识，尽可能地使他不受语言能力的影响。

第五，与一个专题有关的一切项目必须放在一起。这和

一个特殊类型中的一切项目都必须放在一起的建议是不矛盾的。当决定了考查某一专题时，就要决定应当把何种类型的项目应用于这个专题。于是专题就被特殊种类的项目充实起来，它可以避免各类项目之间的重复。有许多应试者无疑地有过接受多重选择和正误问题的客观性考试的经验。多重选择题那一部分所考查的专题也是正误题部分所考查的。结果就对适当的答案提供了许多启示。这种不必要和无用的重复是可以避免的。其方法是用一定种类的项目来考查一个专题并把这些项目放在一起，以期被试者的心向不至于常由处理这个项目到处理次一个项目而发生变化。把有关某一种项目内的特殊方面的一切问题都要问到，并在考试中把它们集合在一起。

第六，考试项目的措词应该适当，以便使学生回答问题的内容而不回答问题的形式。考试项目往往会有对答案来说不应有启示。以致真正不知问题适当答案的被试者根据考试题的措词而得到正确的答案。凡有助于泄露答案的那些字或短语尽量少用（除特殊情况外）。如象“一切”，“常常”“绝对”，“永不”和“不是”等的措词，被试者多半会猜是错误的而非正确的。而含有“某些”“可以”，“有时”“一般”等字的措词则多半是正确的而非错误的。假如这些字或短语在编写项目时无法避免，就必须使其多样化，以便使应试者如果依据形式回答问题会得到不利结果。

第七，一个考试项目不要为另一项目提供线索，而回答某一特殊项目能力也不应依靠回答前一项目的能力。

第八，必须避免引人上当的问题，发现问题中的欺骗性和诡诈性的能力一定要表示被试的学科知识水平上。否则，

这些问题由于使较优秀的学生受骗上当，就可能破坏问题本身的目的。比如“1+1=？”这样的试题出现在层次较高的非数学专业的考试中，就会出现诡诈。

第九，正确的答案必须随机排列。要避免统一的模式。正确答案的次序在被选答案中的百分比不要太大。否则会给应试者猜题得分提供条件，降低效度和信度。

第十，指导语必须完整，举例必须清楚。学生有权知道“主考人”对他的要求是什么，在说明任务时应使他了解和遵循指导语的核心精神。

第七章　分数的衍化与合格分数的拟定

第一节　分数的衍化

每次考试后，主考人或有关部门都要把学生的分数填进表格，再计算一下平均分，以便和其它团体进行比较，只此而已。这评价被试者的成绩未免太简单粗糙。现在大、中、小学每学期都要进行不同学科，不同阶段的考试。由于各科和各个阶段标准不同。因此，同样分数所反映的学习水平不相同。评卷员阅卷后，给予被试者的分数称为“原始分数”，原始分数只代表被试者在这次考试获得的分数。若无其他分数或标准作比较，实难再代表其他意义。考试者通常为了不同的考试目的，把原始分数与其他分数或标准作比较，进行不同的转化。这种方法称为分数的衍化。本节将就分数衍化进行讨论。

一、分数衍化的意义

考试经过设计、命题、预试、实施、评卷等几个环节，得到了每个考生的分数。这并没有告诉考试工作结束。这样的原始分数并没实现考试的目的。考试分数的处理、解释，用以说明一定的问题，从原始分数上是不能体现出来的。

首先一个问题，就是一群考生的成绩分布问题。许多考

试的重要目的之一是了解被考者的知识水平与智能差异的总体状况如何。如平均水平怎样、水平比较接近还是相差很大，优、良、中、及格、不及格的考生各占多大比例，等等。原始分数不可能直接描述团体的基本情况。因此，就需要研究考生分数的分布，必须将分数进行衍化，求得能够说明这种分布的特征值。

第二个问题，是不同考试的分数累计问题。为了提高考试的有效性，常常利用多次、各科考试来实现同一个考试目标，比如高考，要同时考核六、七门课程。学业成绩评定，某学科经常要考试多次。目前，我们常采用百分制累加总分法，这种方法越来越清楚地表现出它的缺点，概括起来有下列问题：其一，它将各科百分制成绩不分青红皂白，一视同仁地相加，这就掩盖了多科之间的主次轻重，忽略了考生要学好每一门课程所必需支付的劳动量的差异，也忽视了考生今后在专业上或工作上相关课程的基本要求。其二，各科百分制成绩累加总分法默许了"一俊遮百丑"或"百俊带一丑"的现象，在挑选人才方面，是挑选面面俱到的全才呢？还是挑选锋芒毕露的偏才？无法区别对待。挑选人才时应当有点倾向性，没有倾向性就会以"量分录用"来代替"量材录用"，这显然是不妥当的。

第三个问题，原始分数对相对测量缺乏含义的直观性．考试可分为绝对测量和相对测量两种。绝对测量所要测试的是多个考生对全部考试内容的掌握程度，它常常规定对全部考试内容掌握的某个百分比（如60%）为及格的标准域。相对测量所要测试的是多个考生间的水平差异或者说每个考生在团体中处于什么地位。由此，相随而产生绝对评分与相对

评分。

绝对评分：分数的高低，决定于考生对考试所要求的全部知识的掌握程度。它提供的信息是考生哪些项目已经掌握，哪些项目尚未全部掌握。它所鉴定的是考生已掌握的知识与预定目标的关系。因此，绝对评分制度通常用于“目标参考性测验”。我们熟悉的百分制和五级分制，就属于绝对评分。

相对评分：分数的高低，决定于考生与旁人比较时，在他所属的考生群（如一个班级、一个年级或高考时报考同一院校，同一专业的考生等等）中所处的地位，它提供的信息是，某考生在这一群体中的名列前茅还是名落孙山。它所鉴定的仅是现有条件下所仅能达到的，其它一概不管。相对评分制度还常用于“常模参考性考试”。

两种评分方法的优劣暂不讨论。问题关键在于，按通常评分的办法，不论是绝对测量还是相对测量，都以完满掌握考试内容为参照，以评分标准为给分依据，进行评分工作。这样评得的分数，对于绝对测量具有直观的意义，对于相对测量则缺乏含义的直观性，必须转换直观意义的分数即相对分数。

二、考试分数的分布

正态分布是统计学上一种常见的概率分布。凡受各种因素影响的随机事件，其出现的概率遵循正态分布规律，考试实质上是对考生掌握的某门课程或某几门课程的全部知识或部分知识和有关智能进行随机抽样测试。考生的考试成绩也是一种受多种因素影响的随机事件，应该基本上符合正态分布。但是，实际考试的成绩是否适合正态分布？属于哪一种

正态分布这还需要进行一些统计研究。

为了进一步了解分数分布的集中趋势和离散程度，还应该求出能够表征这方面特点的集中量数和差异量数。

集中量数有平均数（$\bar{x}=\sum_{i=1}^{n} x_i/n$），中（位）数（就是在按大小顺序排列的一组成绩中，点正中位置的那一数值）和众数（频次最多的分数）。

差异量数有方差、标准差（$S=\frac{\sum_{i=1}^{n} x_i^2}{n}$）、全距（最高分数与最低分数之差），二分位差，四分位差（将分数由小到大排列，按数据个数用三点把分数划作四等分，第二个分点是中位数，第三个分点与第一个分点分数差的一半叫四分差）和平均差（$MD=\frac{\sum_{i=1}^{n}|x_i|}{n}$）

如学业成绩考试，在教与学处于正常情况下，在比较合适的考试命题条件下，一个班级或一个年级的学生考试成绩的分布可按如下给出的比例掌握。

考试成绩分布比例

五级分制		成绩比例%
A	优	不超过15%
B	良	不超过20%
C	中	大约30%
D	及	大约20%
E	不及格	不超过15%

上述给出的指标是接近正态分布的一个范围，只要考试成绩A、B、C、D、E的分布符合上述的要求，而且A与E，B与D基本上保持对称分布，就认为是适合的。当然对于成绩突出班与后进班，考试成绩分布可能超出给定范围。

三、百分比值分

百分比值分是常用的衍化分数的方法，是一个被试者参加考试所得的原始分数与该考试卷满分分数之比，再乘上100就是百分比值分。若一个知识和能力都很高的被试者参加某一考试时，他答对全部试题，获得满分。而知识和能力较差的被试者的分数，只占满分的一个小比率。

例如：某一被试者一门课程考试得70分，试卷的满分为120分，那么说被试者的百分比值分是 $\frac{70}{120}\times 100 = 58.33$。

百分比值分可以只能在同一考试中进行比较。千万不能在不同的考卷所获得的百分比作比较，否则就会不科学。理由如下：其一，不同的考试卷有不同的最高标准。这些是由于不同试卷的试题教学的内容和目标不一致，代表程度不同而产生。因此，不同试卷的满分有不同的意义，而以这些不同意义的满分为基础所衍化的百分比值分，就有不同的意义了。其二，不同的试卷有不同的难度。获得相同百分比值分的被试者对不同难度的试卷有不同的意义。被试者在难度高的一份试卷的得分与较低的另一份试卷的得分比较，是有差别的。为此，不同的考试卷的百分比值分是不可以作比较的，亦不可以相加后作总成绩的比较。

四、位置百分

位置百分是以分数反映出某被试者的成绩在集体中的位置。如对某团体的考试，将团体中数十人的成绩，按好坏排成一长队，就确定了每一成员的百分位置。若给其中某一成员的成绩评了90分，这就表示有90%的人比他成绩低；若评为50分，表示有一半人（50%）的成绩比他低，当然得100分的是第一名，而最后一名为0分。这种分数转化的优点是只要知道某人所得的分数，就知道他在集体中所处的位置，也了解他的水平与集体水平的比较情况。缺点是计算比较烦杂。为了更好地适应这种方法，下面举例说明：

当人数较多，在整理原材料时必须做频数分布，通过频数分布计分表。方法如下：

第一步：做频数分布；

某年级学生跳高成绩累计频数表

组　　限（米）	频　　数	累　计　频　数
0.80—	1	1
0.90—	1	2
1.00—	14	16
1.10—	17	33
1.20—	19	52
1.30—	26	78
1.40—	30	108
1.50—	40	148
1.60—	5	153
1.70—	1	154

n= 154

第二步：求累计频数（如上表）：在频数分布表的右边加"累计频数"栏。将各组频数由上往下累加。如第一组频数是1，累计频数也为1；第二组频数为1，累计频数为1＋1＝2；第三组累计频数为2＋14＝16，最后一组的累计频数就是全部频数n。

第三步：计算百分位置。计算的原理是假设每组内的人数是均匀的分布在这一组限内的，然后找出某一成绩在集体中所处的百分位置，列成计算公式为

$$x\text{成绩的位置百分}=\frac{(x-\text{下组限})\dfrac{\text{组内数}}{\text{组距}}+\text{组下数}}{n}\times 100$$

其中，x为某人的成绩。下组限为某人成绩所在组的下组限。组内数为某人成绩所在组的频数。组下数为某人成绩所在组以下的累计频数。n为总频数。

设有三名被试者的跳高成绩为：1.41米、1.63米、1.12米，试求三名被试者的位置百分。

1.41米的位置百分为：

$$\frac{(1.41-1.40)\dfrac{30}{0.10}+78}{154}\times 100=52.59$$

1.63米的位置百分为：

$$\frac{(1.63-1.60)\dfrac{5}{0.10}+148}{154}\times 100=97.07$$

1.12米的位置百分为：

$$\frac{(1.12-1.10)\dfrac{17}{0.10}+16}{154}\times 100=12.59$$

五、标准分

不同科目的考试分数实际上是通过不同的考卷而得到的读数。上学期数学考试成绩的63分并不等于下学期数学考试成绩的63分，也不能认为语文的80分与数学的80分处于同等水平。

显然不同标准的考试分数是没有可比性和可加性的。为了使被试者各科之间，各阶段之间的考试成绩具有可比性和可加性，比较科学的方法是把被试者的各项考试原始分数都转化为标准分数。然后再进行比较。

美国测验专家W·A·Mecall博士提出用偏差值评分法来评定学生的成绩。由于这种评分法相当于数理统计中随机变量的标准化，所以一般地称这样得到的分数为标准化分数，有时也称作基分数或Z分数。它是以标准差（统计理论上用的表示差异的量数）为单位表示一个分数在团体中所处位置的相对位置数量。Z分数无实际单位，它还可使我们对不同分配的各原始数目之间进行比较。

假设我们对n个被试者，进行了m次测验，X_{ij}是i个被试者在第j次测验中的百分制分数，N_j是第j次测验的团体总评分数，S_j是第j次考试的全班成绩的标准差，则将原始百分数X_{ij}转化为标准化分数Z_{ij}的过程，可按下述步骤进行计算：

（1）计算全体被试者（n），某次（j）考试成绩的总平均分数：

$$N_j=\frac{\sum_{i=1}^{m} x_{ij}}{n}$$

（2）计算全体被试者某次（j）成绩的标准差：

$$S_j=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_{ij}-V_j)^2}$$

（3）计算与X_{ij}相对应的标准分数：

$$Z_{ij}=\frac{X_{ij}-N_j}{S_j}$$

（3）式表明，标准化分数Z_{ij}是以团体平均水平作为比较的参照基准，高于平均成绩给正分，低于平均成绩的给负分。具有正负分数，这是标准化分数与原始百分数区别最大之处。在用百分制总数时，任何分数对总分都有贡献，在标准化分数中却有加有减。这样，两种评价方法得出的结果往往不一致。

例如下表所示，甲、乙两名被试者的原始成绩转化为标准成绩。从原始分数看，两位被试者6科的标准化表：

科目	原始分数（Xij）		总体参数		标准分数（Zij）	
	甲	乙	平均分 N	标准差 S	甲	乙
语文	83	80	82	12.5	0.08	-0.16
数学	77	95	77	5.1	0	3.53
外语	65	61	64	3.12	0.32	-0.96
政治	74	70	71	2	1.5	-0.5
物理	75	76	74.8	1.67	0.12	0.72
化学	75	72.2	74.4	4.62	0.13	-0.48
总计	449	454.2			2.15	2.15

甲的成绩不如乙的成绩。但当甲、乙两人的各科转化为标准化分数后，他们的标准分数的总和都是2.15分，也就是说，按标准分数计算甲、乙两人的Z成绩，他们的水平是一致的。从上表可以观察出：甲的各科成绩均不低于被试团体的平均成绩，而乙只有数学一科考试很好，除数学、物理外，其他各科都在平均成绩之下。因此，出现上述标准分相同是可以理解的。

但是不同考试科目的分数累计时，情况要复杂得多。必须经过如下两个步骤。第一，把原始分数按公式

$$Z_i = \frac{x_i - \overline{x_i}}{S_i}$$

转换为标准分，消除考试难度不同等因素造成分数分布不同所带来的同一分数的“价值”差别，使它们变成为直接可比较的分数。第二，根据不同考试对实现目标的不同作用，给不同考试的分数乘以不同的系数，这个系数通常称为“权重”列成公式为：

$$Z = f_1 Z_1 + f_2 Z_2 + \cdots + f_n Z_n$$

其中n是考试的次数，Z_i是第i次考试某考生的标准分数，f_i是第i次考试的权重，Z是某考生n次考试的累加分数（标准分）。

权重f_i表示的第i次考试的重要程度。它的确定通常是采用经验法与多元回时法。

标准分数虽可以直接反映该成绩在团体中的地位，在科学研究中常用以综合分析问题。但由于它有正有负，使用不

方便，因此，可以将标准分数变为标准百分。在计算中常用公式：

$$Z_i=\frac{x_i-\bar{x}}{S}\times 10+60$$来计算标准分数

其中，$\bar{x}$表示某科的平均成绩，S表示某科的标准差。X_i第i个被试者的某科的原始成绩，Z_i表示第i个被试者的某科标准成绩。

标准分是建立在原始分数服从正态分布基础之上的。如果原始分数不服从正态分布，标准化后的标准分仍不服从正态分布，这是可从标准化分数的计算公式看出的。如果原始分数不服从正态分布，而是正偏态或负偏态，那么，标准分也为正偏态或负偏态，它弥补百分制分数在一般情况下缩小了高水平学生之间的差距及低水平被试者之间差距等弱点的优越性不能得以充分的体现。

然而，无论什么分布，中位数两侧分布的考生比例是相等的，各占50%，因此，我们可以中位数为基准点建立一种标准分。称为中位数标准分。

中位数标准分的定义：$MZ=\frac{x_i-\beta}{\alpha}$

其中β为中位数，X_i为第i个被试者的原始分数。n为被试者分数。X为中位数标准差。

$$\alpha=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\beta)^2}$$

原始分数化成中位数标准分以后，解决了当原始分数不服从正态分布，标准分也不服从正态分布的弱点。中位数标准分比标准分更优越，更有较强的可比性和可加性。其原因

主要是改变了原始参照点，使各科具有相同的基准点——中位数“零点”，相同的度量单位——中位数标准差（单位1），并且，各科中位数标准分的中位数两侧分布的被试者数量相同。

第二节　选择题评分问题的研究*

选择题的解答包含有随意猜测的因素：一是答对的题中有不懂而猜对的因素；二是答错的题中有能答对的因素；三是没有做完的题或未做的题中既有可能答对的因素，也有答错的因素。但是这不能否认客观式选择题的时代作用和意义。若不知正确答案的考生完全凭猜测答题，则答错对机会服从于《概率原则》。本节试图从概率论的角度出发就客观式选择题的评分问题进行探讨。

一、猜测分数的矫正

运用概率原则可决断选择题的可信性范围。当某事物随机出现的概率 α 较小时（通常 $\alpha < 0.005$），重复这个随机事件的次数不太多时，这时称该事件为小概率事件。可以认为小概率事件几乎是不可能发生的。若选择题答对的概率 α 在某一个规定数以内，便可排除随机猜测的可能。答对的题数具有可信性。然而，若干题中要答对多少题才能排除猜测因素呢？分析如下：

*徐玖平《客观式选择题评分问题的初探》，教育测量与标准化考试研讨会。

对一组t道n选1的选择题，运用二项式展开得做对t_0题或t_0题以上的积累概率为$g_t^n(t_0)$

$$g_t^n(t_0)=\left(\frac{1}{n}\right)^t+C_n^{t-1}\left(\frac{1}{n}\right)^{t-1}\left(\frac{n-1}{n}\right)+C_n^{t-2}\left(\frac{1}{n}\right)^{t-2}\left(\frac{n-1}{n}\right)^2+\cdots\cdots+C_n^{t-t_0}\left(\frac{1}{n}\right)^{t-t_0}\left(\frac{n-1}{n}\right)^{t_0}$$

假设小概率事件的概率为α_0，那么满足$g(t_0)<\alpha_0$所对应最小t_0就是所要求的做对题的数目。即在t题中做对t_0题或t_0题以上就可能排除考生的猜测因素。

对考生的分数应怎样看待，笔者认为：由于有猜测因素，要进行修正分数。怎样修正？这需要应用概率论的知识。假设某考卷每题有n个答案供选择（只有一个正确），其评分标准是：选对得a分，选错扣b分。若有人凭运气猜题，随机地选择答案，则猜对的概率是$\frac{1}{n}$，猜错的概率为$\frac{n-1}{n}$，若以随机变量ξ记每题目的得分数，则其概率分布为：

ξ	a	b
p	$\frac{1}{n}$	$\frac{n-1}{n}$

故数学期望$E(\xi)=a\cdot\frac{1}{n}-b\cdot\frac{n-1}{n}$

若$E(\xi)\leqslant 0$意味着猜题失败。取最低的失败（$E(\xi)=0$）来考虑，则有：

$a-b(n-1)=0$　　从而$b=\frac{a}{n-1}$

一般情况b、a均为正整数，所以$\begin{cases}a=k(n-1)\\ b=k\end{cases}\quad(k\in Z)$

这就是合理的得、扣分比例标准。

根据上述基本思想，下面给出修正分数计算公式：

$$S=R+\beta\cdot\frac{Sng(\alpha_0-g(t_0))-1}{2}\cdot\frac{t_1}{n-1}\cdot m+\frac{t}{n-1}\cdot m \quad (A)$$

这里 $sng(\alpha_0-g(t_0))=\begin{cases}1 & \alpha_0-g(t_0)>0\\ 0 & \alpha_0-g(t_0)=0\\ -1 & \alpha_0-g(t_0)<0\end{cases}$

R是考生实得分数，S是修正分数，m是每题应得分数，t_1是答错题目个数，β是一个常数。

上公式表明，若考生做对t_0或t_0题以上，就不扣分，若考生做对题数小于t_0，那么就要扣分。

举例：

10道4选1的选择题，做对几道才能排除猜测的可能呢？运用二项式展开知识及上述判断原则可证明只有做对b题或b题以上才能排除其成绩出于猜测因素。叙述如下：

建立$(1/4+3/4)^{10}$的项式，其中1／4是每题的正确概率，3／4是错误概率。指数10是总数，学生的各种答卷情况可归于11种组合，每种组合出现的概率恰等于$(1/4+3/4)^{10}$展开式的11项数。

根据教育统计学概率原则，概率在0.05以下，称为小概率事件，由上表看出：要$g^4_{10}(t_0)<0.05$则$t_0\geq 6$，用（A）式计算得考生修正分数。下表所示。

用（A）式计算时，若$\beta>0$修正分数实际上是分数增值了。这可用于资格、水平考试等。对于学校班级中期或期

回答情况	概　率	实得分数	修正分数
全对	10^{-6}	30	40
9对1错	0.000416	27	37
8对2错	0.003486	24	34
7对3错	0.003506	21	31
6对4错	0.019728	18	28
5对5错	0.078127	15	25 - 5/2
4对6错	0.224125	12	22 - 6/2
3对7错	0.474407	9	19 - 7/2
2对8错	0.755975	6	16 - 8/2
1对9错	0.943687	3	13 - 9/2
全错	0.971689	0	10 - 10/2

末考试可取 $\beta<0$，但要适当控制。原则上不能使被加分数的考生的分数高于在该考生前的考生的得分。

（若对（A）的数学理解有困难的可用

$$S=R\pm\frac{t_0}{N-1}m+\frac{t}{N-1}m$$

（B）这里的加减在同一集体中必须一致）

二“轮盘赌”的思考

用选择支数为n的客观式选择题施考时，对考生来讲，他可以把n个答案看成各有多少是正确的因素。也就是说，占正确答案的百分比是多少。若考生水平高，那他就会认为正确那个选择支数是百分之百的。因此，把该题的所有分数都押在那个选择支数上。若考生成绩平常。那他就不会很准确选中正确答案，而他可以将该题分数分散押在他认为正确因素的那些支数上。这样一来各种情况的考生都能比较客观地反映成绩。那么怎样计算这些分数呢？此问题实际上是一个二元相关“一致性”的问题。如何从数量上表示正确答案与考生的答案“一致”的程度？这不能用平均的办法来加减，这里我们引用热力学中的“相对熵”的概念，来解决这个问题，下面给出一计算公式：

$$h_i = m_i \cdot \left[\frac{\ln(g+1)}{\ln(m_i+1)}\right]^c$$

h_i表示第i题的得分，g表示考生在正确的选择支上的投放分数，m_i表示i题的满分。c是常系数，它可以系统而有效地控制加分和扣分的幅度。

应用举例：取m＝2、n＝4、c＝1，下面是甲、乙、丙、丁四个考生对某一题的回答情况：

选择支数（n=4）	A	B*	C	D
得分正确分布	0	2	0	0
甲考生答题情况	0.5	1.2	0	0.3
乙考生答题情况	0	2	0	0
丙考生答题情况	2	0	0	0
丁考生答题情况	0.5	0.5	0.5	0.5

B*表示正确的答案

$$h_{乙}=m_i\cdot\left[\frac{\ln(g+1)}{\ln(m_i+1)}\right]^1=2\cdot\frac{\ln(2+1)}{\ln(2+1)}=2$$

$h_{甲}=1.43$，$h_{丙}=0$，$h_{丁}=0.73$

上例可看出，全错得零分，全对得满分。部分对得部分分。

三、时间参数

时间是一切事物演化史上的最客观的量度。以时间作为参量去研究事物的性质和变化，是自然科学和社会科学等各部门中重要方法和手段。对于客观选择题评分问题也同样如此。比如，某考生10分钟答完5题和另一考生15分钟答完5题对错一样，成绩应该谁高些？当然是前者。但是，如果两名考生所回答全部考题时间不一样，对错也不一样，那么问题就比较复杂了。针对上述问题，把考生的成绩与考生答卷的时间对照联系起来，确认考生的分数是实得分数x与完成答

卷的时间y的二元函数，是比较客观的作法。为此有：G=g（x、y），G是标准分数。

$$G = x\left[1+\frac{e^{Y}-1}{ke^{T}}\left(\frac{x}{H}\right)^{k'}\right] \quad (c)$$

x表示考生实得分数，y表示答卷所用的时间，规定“准时”交卷的时间为零。提前为正，延后为负。k、k′均为正常系数。H为试卷满分。k、k′的取值可由经验而取得，也可以于事前举行模拟考试，以求由实验结果通过数学计算而得到理想的最佳数值。

上述（C）式表明，1° 若考生“准时”交卷则y=0，那么G=x。说明考生在智力的速度反应能力上一般，没有什么特别的能力，因此不加分。也不扣分。

2° 若考生提前交卷，则T＞0，说明该生是智力高或具有快速反应能力，此种人才当然要比“准时”交卷的考生要在某些方面强些。因此，应该加分。∵$e^{y}-1>0$，∴加分为

$$x\cdot\frac{e^{Y}-1}{ke^{T}}\left(\frac{x}{H}\right)^{k'}$$

3° 若某考生延后交卷，则T＜0说明该考生思路发挥较慢，反应迟钝。因此，该扣分。∵$e^{y}-1<0$，∴扣分为

$$x\cdot\frac{e^{Y}-1}{ke^{T}}\left(\frac{x}{H}\right)^{k'}$$

加分与扣分都是与原来所得分有关系。这不可能出现白卷得分的。由于e^{x}的图象不是直线。加分与扣分的幅度不会是直线增减，而是曲线逼近的。

举例：甲、乙、丙、丁、戊五考生的成绩修正分数：

k=1　　k′=1　　T=2　　H=100

考　生	时　间	实 得 分	修 正 分
甲	1	80	94.8
乙	-1	100	36.8
丙	0.5	90	97.1
丁	0.1	70	70.7
戊	0	60	60

以上所提出的公式，就其计算，比较麻烦。对于较烦琐的计算问题，可采用电子计算机来代替人工操作。这是可行的。当然，就公式计算的本身是需要实践的验证的。

第三节　合格分数的拟定

合格分数是反映被试者的能力和知识的一个度量标准。被试者每参加一次考试所获得的分数，要反应被试者对所考学科的学习能力。它是划线的分数，被试者的考分高于合格分数，则视被试者的能力高，反之，亦然。换一句话讲，被试者的分数在这条界线以上，表示他达到合格以上的水准；在这界线以下的分数，表示他未能达到合格的水准。而相等于这条界线的分数则表示获得这些分数的被试者则刚好达到合格的水准，是一组“边界”的被试者。

要划分数线区分被试者的“合格”与“不合格”两组，通常是比较困难的。原因是由于“合格”与“不合格”是二

值判断，造成被试者在考试中不能正确反映他的真实能力。一次性的考试分数并不代表他真正能力应所有的分数，因而出现考试分数与应有分数的差别，这一类是测量误差。另一类是分组误差，由于施考者只以划线分数来区分合格与不合格两组，从而把能力高而考试成绩低的被试者错误地划入不合格组；或是把能力低而测验成绩高的被试者错误地划入合格组。考试误差是客观存在的，无论使用任何方法，在目前的条件下亦不能把它全部消除，只可把它的程度减弱。

一个合格分数，若以常模为准则，施考者需在考试后计算学生的考试分数，然后定下合格人数的比例。再从这个百分比值转化为一个划线分数。若以试题数目所代表能力高低为准则，施考者不需要考虑被试者在测验时的表现，只需根据他认为达到最低水准的能力，转化为考试分数，定出划线分数。不论应用什么方法来拟定合格分数，决策人是施考者本身。合格分数是个人的决定，容易产生主观错误。那么，如何拟定合格分数呢？本节将介绍两类常用的方法。

一、聂刁斯基（Nedelski）评核法：

在1954年，测量学专家聂刁斯基提出：一个客观性考试的合格分数是由有关专家评审测验试题的内容来决定的。它必须依循四个步骤：

第一，选择专家，评审试题的人员必须有足够的能力判断考试卷内各试题所考核的知识和能力的水准的专家，这项工作比较困难，常常只能选择部分的有关专家。

第二，鉴定“边界”知识和能力。将选出来的专家们组织起来，对试卷所考核的知识和能力的水准进行讨论，并理

出可接受的水准和不可接受的水准及两者之间的水准，这个两者之间的水准是代表“边界”知识和能力。

第三，专家评分，各专家需根据每一试题的“边界”知识和能力，自行计算具有“边界”知识和能力的被试者在各试题的得分。然后把各试题的得分加起来，所求得的总和就是个别专家所拟订的合格分数。

第四，综合专家的意见，把各专家拟订的临时合格分数，计算平均值或中位值“剪修”平均值，得出合格分数，换句话说：平均值是临时合格分数总和与专家人数之比，中位值是从成绩高低排列临时合格分数选取中间位置的临时合格分数。“剪修”平均值是去除最高值和最低值后的临时合格分数的平均值。

下面是计算合格分数的平均值，中位值和“平均值的例子”：

临时合格分数的人数	临时合格分数	除去最高和最低分的临时合格分数
1	97.3	—
2	86.2	86.2
3	75	75
4	74	74
5	61	—

总分＝393.5　　　　总分＝235.2

平均值$=\frac{393.5}{5}=78.7$　　　　“剪修”平均值$=\frac{235.2}{3}=78.4$

中位值＝第1位值的临时分数＝67.00

在一九七一年，测量学家安国辅将聂刁斯基法稍作修改，提出：只需专家决定“边界”被试者对每一试题的机率，不判断“边界”学生对每一试题的误差考察的辨认能力。因此，这个方法不仅适用于各项选择题的考试，而且还适用于其他类型的考试。安国辅与聂刁斯基拟定合格分数的四个步骤大致相同。前者在第三个步骤计算每题的合格分数时，计算“边界”学生答对每题的机会率，后者则计算“边界”学生在每题难以辨认的选择考察的数目，再求倒数。以下是聂刁斯基方法与安国辅方法的两个例子，予作比较。

试题编号	“边界”学生不能正确辨认选择答案的数目	“边界”学生答对的机会率	用聂刁斯基法“边界”被试者的得分
1	3	0.95	1/3=0.33
2	4	0.20	1/4=0.25
3	1	0.90	1/1=1.00
4	5	0.60	1/5=0.20
5	2	0.75	1/2=0.50
6	1	0.40	1/1=1.00
7	3	0.50	1/3=0.33
8	4	0.20	1/4=0.25
9	5	0.25	1/5=0.20
10	2	0.35	1/2=0.50
11	1	0.45	1/1=1.00
12	1	0.25	1/1=1.00
13	4	0.40	1/4=0.25
14	3	0.75	1/3=0.33
15	3	0.90	1/3=0.33
		总分=7.85	总分=7.47

应用聂刁斯基方法计算的专家个人所拟订的合格分数＝7.47，应用安国辅方法计算专家个人所拟定的合格分数 7.85

在一九七二年，测量学家伊实尔提出：每位专家需把试题分类，填入难度与重要性的二维向度表内，并决定“边界”被试者答案对表内每一小格的试题的百分比，然后把各题的这个百分比转化为小数位数值，再把这些小数位数值加起来，就是该位专家所拟定的考试合格分数。应用伊实尔法计算合格分数的步骤与聂刁斯基法大致相同，在第三个步骤有差异，前者需要先制订试题难度与重要性的二维向度表，然后根据这二维的向度计算每题的合格分数，以下是应用伊实尔法把试题分类例子：

难度 / 试题编号（“边界”学生答对试题的百分比）/ 重要性	容易	适中	困难
非常重要	1、17、18（0.75）	2、15（0.75）	3、8、19（0.16）
比较重要	9.25（0.65）	7、14、16（0.35）	4、20（0.42）
重　要	—	6、13、24（0.71）	5、21、22（0.91）
一　般	11、12、23（0.50）	—	10（0.20）

应用伊实尔法计算合格分数的例子

试题类别	"边界"学生答对试题的百分比	同类的试题题目	合格分数
非常重要及容易	0.75	3	0.75×3=2.25
非常重要及适中	0.75	2	0.75×2=1.5
非常重要及困难	0.16	3	0.16×3=0.48
比较重要及容易	0.65	2	0.65×2=1.30
比较重要及适中	0.35	3	0.35×3=1.05
比较重要及困难	0.42	2	0.42×2=0.84
重要及容易	—	—	—
重要及适中	0.71	3	0.71×3=2.13
重要及困难	0.91	3	0.91×3=2.73
一般及容易	0.50	3	0.50×3=1.50
一般及适中	—	—	—
一般及困难	0.20	1	0.20×1=0.20
专家个人拟订的合格分数=13.98			

上述拟订合格分数的办法，有一个弱点，它离不开专家的评审试题和拟订合格分数。如一个是学校的一般期中或期末考试，要拟订合格分数，我们可选取有教学经验的教师担任专家角色。当然，这些"专家们"必须对考试所考核的内容和能力有足够的认识。

二、"边界组"评核法：

一九八0年，测量学专家黎贵思敦（luincston）提出：从"边界"被试者考试所得的分数，计算出合格分数的新方法。其优点是不单从测验试题的评审中，订出合格分数。这

样节省了人力、物力。主要步骤如下：

第一，选择评审员，评审员必须具有忠实和客观表达自己的意见的品质，必须是对应试的被试者的知识和能力有充分认识。评审员的数目越多与拟定合格分数的准确性，客观性成正比。

第二，鉴定考卷所考核的内容和能力，分析被试者的答卷的情况，评细分三个类别：合水准（可接受水准），不合水准（不满意或不可接受的水准）。"边界"水准。

第三，辨认"边界"被试者，获取他们的考试分数，以作下一步统计。

第四，评审员拟订合格分数。从"边界"被试者的考分中，计算中位值。中位值不受过高和过低分数的影响，较能代表他们的成绩。因此，中位值就是评审员拟订的合格分数。若"边界"被试者的分数相近，中位值的代表性高，反之，若他们的分数相差大，则中位值的代表性低。

第五，综合评审员的意见，用聂刁斯基法第四个步骤所用的方法，计算平均值或中位值"剪修"平均值。最后得合格分数。

举例如下：

应用“边界组”评核法计算合格分数

“边界”被试者名次	“边界”被试者考试分数
1	79 （最高分）
2	73
3	70
4	68 （高中位值）
5	67 （低中位值）
6	54
7	51
8	50
9	49
10	47 （最低分）

$$中位值=\frac{68+67}{2}=67.5$$

平均值＝60.8

剪修值＝60.25

评审员所拟定的合格分数＝67.5

上述的方法从形式上看，似乎比较合理。但是，仔细考虑，问题解决的不是想象的那样完善。比如，不是“边界”被试者的人，是否一定是合水准的呢？回想一下，“边界组”的评核方法却不是那样。黎贵思敦在提出“边界组”评核法的同时，还提出“对比组”评核法，该方法是让评审员根据被试者的知识和能力，把他们分为合水准与不合水准两组。对两组被试者的考试分数都需用作计算合格分数。对人数过多的情况，评审员可抽取有代表性的被试者的样本作分析。然后制定考试分数合格者与被试者所占的比例。根据考试分

数和被试者的频数分布情况，排出一个分数水平序列。列出隶属每一分数水平的合水准组与不合水准组学生人数。计算出合水准所占的百分比。最后采用“分数平滑法”，调整合水准组学生在每一考试分数水平所占的比例，给出各评审员个人合格分数，再按聂刁斯基法的第四个步骤，拟评出合格分数。

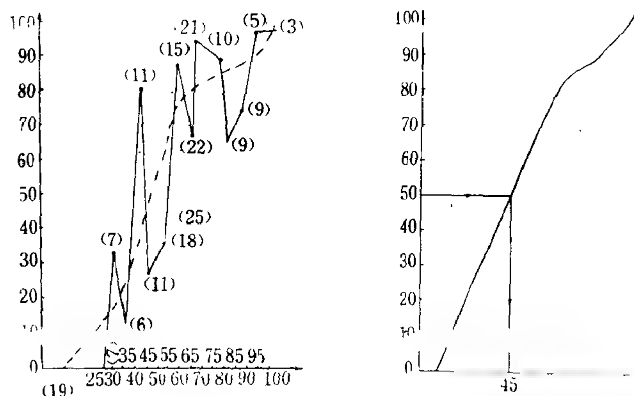
“分数平滑法”是把不规则的合水准组被试者在各个考试分数水平所占的百分比作调整，使这些百分比成为规则的序列。“分数平滑法”有很多种技巧，通常的有两种：一为“曲线评滑法”，二为“移动平均值法”。

“曲线评滑法”是先划出座标，合水准组学生的百分比为纵座标，考试分数水平的中间值为横座标。考试分数水平的中间值是代表同一考试水平的各个考试分数。然后把隶属每一考试水平的合水准组学生的百分比以“点”在“座标”图显示出来。这些点通常在座标图上是不规则的离散点。我们用一平滑的曲线划在座标图上，以代表所有点的平均位置。下面是用“分数平滑法”拟订合格分数的例子。

设某被试的成员为197人，经过评审员们的分析把他们分成不合水准与合水准的两个组，然后计算合水准组被试者所占百分比，下表是某一评审员A的分析。

考试分数水平	被试者人数			合水准组所占百分比
	合水准组	不合水准组	总人数	
96—100	3	0	3	100
91—95	5	0	5	100
86—90	7	2	9	77
81—85	6	3	9	66
76—80	9	1	10	90
71—75	20	1	21	95
66—70	15	7	22	68
61—65	13	2	15	86
56—60	10	15	25	40
51—55	7	11	18	38
46—50	3	8	11	27
41—45	9	2	11	81
36—40	1	5	6	16
31—35	2	4	6	33
26—30	0	7	7	0
0—25	0	19	19	0

应用“平滑曲线”法调整合水准组所占百分比：



假若评审员A拟定的合格百分比为百分之五十（一般而言，以接近百分之五十作为合格百分比是适合的），这表示达到这个合格标准的合水准与不合水准的被试者各占一半，评审员也可根据不同的考试目的和不同的被试者特性，把合格百分比高低）的及格百分比，作一条与横轴平行的直线。交平滑曲线于一点，再从这点作垂直横轴的垂线。交横轴于一点，该点横座标为45，这就是该评审员A拟评的合格分数。

其他的评审员的合格分数分别为47、60、48、34、52、56。

合格分数的拟定

临时合格分数的名次	临时合格分数	除去最高、最低分的临时合格分数
1（最高）	60	—
2	56	56
3	52	52
4	48	48
5	47	47
6	45	45
7（最低）	34	—
总分=342		总分=248

平均值$=\frac{342}{7}=48.85$

“剪修”平均值$=\frac{248}{5}=49.6$

中位值=第四位值的临时分数=48

“移动平均值法”是合水准组被试者与不合水准分出来后，以每三个先后次序相连的考试分数水平来计算隶属这三个水平的合水准组中被试者所占的百分比。当然这个百分比虽然从隶属这三个水平的合水准组被试者与隶属相同的三个水平的合水准两组被试者的比例求得，它是代表隶属于这三个考试分数水平的中间一个水平的合水准组学生在这个中间水平所占的百分比。照用上述例子，用“移动平均值法”来求合格分数如下：

考试分数水平	学生人数		“平滑”合水准组被试者所占百分比
	合水准组	两组总人数	
96—100	3	3	-
91—95	5	5	$100\times\frac{3+5+7}{3+5+9}=88$
86—90	7	9	$100\times\frac{5+6+7}{5+9+9}=78$
81—85	6	9	$100\times\frac{7+6+9}{9+9+10}=78$
76—80	9	10	$100\times\frac{6+9+20}{9+10+21}=87$
71—75	20	21	$100\times\frac{9+20+15}{10+21+22}=83$
66—70	15	22	$100\times\frac{20+15+13}{21+22+15}=82$
61—60	13	15	$100\times\frac{15+13+10}{22+15+25}=61$
56—60	10	25	$100\times\frac{13+10+7}{15+25+18}=51$
51—55	7	18	$100\times\frac{10+7+3}{25+18+11}=37$
46—50	3	11	$100\times\frac{7+3+9}{18+11+11}=47$
41—55	9	11	$100\times(\frac{11+11+6}{3+9+1})^{-1}=46$
36—40	1	6	$100\times\frac{9+1+2}{11+6+6}=52$
31—35	2	6	$100\times\frac{1+2+0}{6+6+7}=15$
26—30	0	7	$100\times\frac{2+0+0}{19+7+6}=6$
0—25	0	19	——

若选取百分之五十一的一个百分比为合格百分比，找出这个百分比所需的考试分数水平，用聂刁斯基法的第四步得：

应用“移动平均值”方法计算合格分数：

考试分数水平	“平滑”后的合水准组被试者占百分比
—	—
61—65	61
56—60	51
51—55	37
—	—

合格百分比＝51%

相对合格百分比的考试分数水平＝56至60分

56至60分的中间值＝58分

拟定合格分数＝63分

第八章　试卷分析与研究

第一节　试卷分析

一、试卷分析的意义

考试实施之后，对试卷进行分析是考试工作的一个重要环节。可以讲，没有试卷的分析，就等于该次考试没有结束，或是不完整的考试。因此，它的意义很大，工作的技术性很强。为什么要进行试卷分析呢?

首先，是评价考试的质量的需要。

考试是一个系统，它必然有评价这一子系统，这一环不可缺少，否则许多工作就无法做了。比如，考试工作的质量与实际效果，目标达到标准没有，命题工作中得与失，成绩的可靠性大不大、效度高不高等等有关系。对于主考单位来说，这些都是必须解决的问题，工作十分有意义，很有探讨的必要。

当然，评价考试质量的因素与方法是多种多样的。如象检查已经做过的工作，查其疏漏，对考生进行跟踪调查、研究、了解考试成绩反映的情况与考生后来情况的相关程度，等等。但是，我们必须注意考试工作自身的某些特点，特别是，由于考试试卷提供的材料详尽、具体，除可供定性分析处理外，还可以进行定量分析，从而得出比其他方法更精确的评价。为此，试卷分析就成为评价考试质量的重要、有效的方法。

其次，有利于充分发掘考试中储存的信息。

试卷分析，是对试卷中储存的信息的进一步发掘和利用，这信息不只对检查和改进考试工作有重要的价值，对于改进对考生的教学工作也有重要的价值。如对不同地区，不同学校，不同班级的考生试卷进行对比，分析教学质量的差异，分析考生答卷中的典型事例，找出其中带有普遍的问题，对先后举行的同类考试的试卷进行对比研究，分析考生的具体进步和存在的不足等等。对于促进教改，提高教育质量是大有益处的。

再次，有利于考试工作的改进。

对于将要开展的同类考试来说，进行了一次最真实的预试。由于同类考试的考试内容和标准相对稳定，考试的设计蓝图大致相同。因此，它具有很大的可比性，能够为改进考试设计工作，使之更符合考生实际提供的重要信息。能够为改进命题工作提供大量的信息，提高试题和试卷的编制质量。同时，对于题库的建立、提高和改善来说，它就是一次理想的预测。试卷分析对于改进试题和试卷编制工作的作用将更直接，价值也更大。是对印卷、评卷等项工作的一次检查，有利于提高考试组织，管理的工作质量。

二、试卷的定性分析

对任何事物的分析方法，都可从定性分析与定量分析入手，两者互相补充，相互结合。分析试卷也不例外，一般采用定性分析法与定量分析法。分析试卷和试题的质量，以定量分析为主，定性分析为辅；分析考生解答中的具体问题，以定性分析为主，定量分析为辅。

定性分析中，要辅以定量的方法，进行必要的统计分析。分析过程中要注意记录与资料保存，分析之后要分专题写出分析报告。后面将给出实例，这里不多谈。当然，定性分析的方法，远不至于此，通常需要同时对考生和教师的调查、结合进行分析。如通过召开座谈会，听取考生和教师对各种试题的看法，了解考生对某些问题的解答中的具体思路与做法，收集社会的反响，等等。

如要了解考生答卷所反映出的教学中存在的问题，可根据统计分析提供各道试题的难度资料。从优、良、中、及格、差考生中分层抽取若干份试卷，对难度较大亦即考生普遍没有答好的试题进行具体分析，找出考生解答中带有规律性的错误；要研究学习比较差的学生的问题，可抽取若干分数较低的试卷，将试题按考核的具体内容分类，分析考生对各类问题的解答情况，找出普遍性的问题；要了解考试成绩异常的考生的原因，则可选取出该考生的试卷，逐题进行分析；要了解不同地区考生的具体差异，可按同一比例分别从优、良、中、及格、差考生中抽取若干试卷，对考生的解答进行比较分析。

三、统计分析

统计分析是试卷分析的主要方面，为研究试卷和试题质量而进行的统计分析，主要包括以下项目。

①考试的信度分析

②考试的效度分析

③考试的难度与区分度分析

④选择题各误道答案的考生选择情况分析

⑤考试成绩分布的统计分析

为进行上述分析，通常采取下述的工作步骤：

①抽样

②数据整理——造表登记

③绘制频数分布曲线——考试曲线

④计算难度

⑤计算效度

⑥计算信度

⑦计算区分度

⑧写出分析报告

考试的信度、效度、区分度和难度在第三章分别作了介绍，选择题的误道，多选答案在第七章第二节作了专题研究，这里只作工作步骤介绍。

四、实例及其简单的讨论。

试卷的统计分析，其目的：一是评价试卷的质量，二是获得考试反馈的各种信息；三是有利于考试工作的改进。下面将举某市的一次高中入学考试的数学试卷的统计分析为实例来说明试卷统计分析的一般方法。

（一）抽样

规模较大的考试，试卷量很大。若全体作为统计分析的对象，工作量太大。因此，一般都在全体中随机地取出部分——样本，用样本的统计分析来推断全体的情况。在试卷抽样中要遵循以下三条原则：第一是随机原则，使抽得的样本，能够代表总体；其次是可行性原则，即抽样的方法在实际中是可行的；最后是信息性原则，即所抽的样本尽可能反

映出分析时所期望的各种信息。因此，在抽样前必须综合考虑这三条原则，从而设计抽样方案，使获得最令人满意的样本。

一般试卷抽样的方法[1]

1、可按考生在各个分数段的比例进行分层随机抽样。这种方法抽得的样本具有典型性。但它必须先要作好总体分数段的频数统计，无疑工作量很大，故一般不予采用。

2、按考生的编号进行机械抽样。

3、按学校类型的考生的比例进行分层整群抽样。

某市的“中考”抽样采用的是最后一种方法。这是因为，首先是市里需要获取大面积的普通中学与职业中学的教学信息。这两类学校的考生分别是2627和2694人，各占总考生5705人的46.1%与47.1%，其比例基本上是1：1。因此按1：1对两类学校的考生进行分层抽样。这样所得的样本，虽没有包括重点中学的考生，但由于重点中学的学生已大部分直升高中，参加“中考”的考生仅384人，占总数6.7%，比例数小，对试卷质量仅有比较小的影响，这一点在分析中做到心中有数就可以了。

其次市招办规定只能整本（每本试卷25份）查阅。因此就采用分层整群抽样。抽样的另一个问题是确定样本的容量n，确定n的方法是两种：一种是当总体很大时，就取n= 370。理由是在统计分析时，要用到样本的27%为高（低）分组，这样它的容量正好是100，便于计算。

另一种是通过计算来确定n。某市就采用了计算的方法。

---

①参见：徐惠仁、吕听听著《怎样进行试卷的统计分析》

具体做法是：

先确定n的大概范围。按统计上的要求，容量上千的总体，一般按$\frac{1}{30}$～$\frac{1}{20}$取样，这样对考生为5705的总体所取的样本，n可为：$5705\times\frac{1}{30}\sim5705\times\frac{1}{20}$，即在170－250之间。再因为是按1：1分层整群抽样的，所以在普通中学的88本试卷中随机抽取5本其中一本作备用，在职业中学的90本试卷中也随机抽取5本（一本作备用）。这10本要求分布在全市10所或8所不同的中学，若有重复，再重行随机抽取。这样就暂时取得了8本计192份试卷的一个样本。然后计算这$n'=192$样本，得分的标准差$S'=17.6$。统计上计算样本容量n的计算公式是$n=1+\frac{4s'^2}{\triangle^2}$。

其中△为精度，即为均数区间估计中区间长度的一半。

即$\triangle=1.96\times\frac{s'}{\sqrt{n'}}$（信度95％）

代入上式并化简得

$n=1+(\frac{2}{1.96})^2n'=1+(\frac{2}{1.96})^2\times192\approx201$。

为了计算方便就取$n=200$。

最后尚需补8份试卷，即在两本备用卷中以随机抽样的方法，各取4份。

到此我们就取得了要分析的一个容量为200的样本。

（二）数据整理——造表登记

1、造表（见附表）。

2、登分统计：按高分到低分顺序登记（见附表），前

54人为高分组，最后54人为低分组（即27%）。

3、计算样本的平均分$\overline{X}=53.2$，标准差$S=17.79$，连同样本容量$n=200$，三项指标写在表的右上角。

（三）绘制频数分布曲线——考试曲线

1、按分数段列出频数分布统计表。

2、绘制频数分布曲线。

3、检验分布曲线的正态性。这里用X^2的拟合性检验。具体进行如下：

（1）把分数分成如下五个段：$\overline{X}-2.5S\sim\overline{X}-1.5S$；$\overline{X}-1.5S\sim\overline{X}-0.5S$；$\overline{X}-0.5S\sim\overline{X}+0.5S$；$\overline{X}+0.5S\sim\overline{X}+1.5S$；$\overline{X}+1.5S\sim\overline{X}+2.5S$。若分布呈正态的，那么这五个分数段中分数的频数的比例应是70%：24%：38%：24%：7%。

因为，$\overline{X}=53.2$，$S=17.79$，

所以这五个分数段应是

9～26，27～44，45～62，63～80，81～98。（分成六个、七个都行，而且越多越精确）

（2）统计这五个分数段的实际频数f_i，并计算这五个分数段的理论频数f'_i。制成下面的列联表。

f_i	9～26 14	27～44 51	45～62 69	63～80 53	81～98 13	Σ	差异
f'_i	14 200×7%	48 200×24%	76 200×38%	48 200×24%	14 200×14%		

$$X^2=\frac{(14-14)^2}{14}+\frac{(51-48)^2}{48}+\cdots\cdots+\frac{(13-14)^2}{14}=1.42$$

而$X^2_{0.1}(4)=7.78$，$\therefore p>0.1$（接受无差异假设）

因此得频数分布曲线是呈正态性的。

（四）计算难度（得分率）

1、计算每题的难度pl，27%的高分组的难度P_H,27%的低分组的难度P_L，及试卷的难度P（各题难度的平均值）。数据见附表。

2、难度指标

0.7以上的为较易题；

0.3—0.7之间的，为中等难度题；

0.3以下的，为较难题或难题。

此考卷的八道题中$p_1=0.34$，$p_2=0.41$，$p_5=0.34$，$p_6=0.355$，即1、2、5、6四道题属中等难度题，占50%。而$p_3=0.26$，$p_4=0.27$，3、4两道题属于较易的，占25%。$p_7=0.80$，$p_8=0.95$。即7、8两道题属较难题与难题占25%。难易分布较好，且试卷难度$p=0.488$，是中等难度。因此从难度来看，此试卷是好或较好的水平。

（五）计算区分度

1、计算每题的区分度（见附表区分度栏）

计算区分度的方法除了在本书中已经介绍的费拉南根查表法外，这里再介绍两种方法：

一种是计算点二列相关系数r_b，因为区分度可以看作是该题的得分与不得分的点二列相关问题。因此，

$$r_b=\frac{\overline{X}_p-\overline{X}_q}{S}\sqrt{pq},$$

其中$\overline{X}_p$是答对该题之考生卷面成绩的平均分，

$\overline{X}_q$是未答对该题之考生卷面成绩的平均分（得题分一半以上的为答对，否则为未答对），

S是试卷得分的标准差，

P是答对该题之考生对总数的比例，

q是未答对该题之考生对总数的比例。例如此试卷的一（3）答对的是48人，$\therefore p=\frac{48}{200}=0.24$。答错的是152人，$q=\frac{152}{200}=0.76$且统计得$\overline{X}_p=69.31$，$\overline{X}_q=48.36$，$S=17.79$

所以$r_b=\frac{69.31-48.39}{17.79}\sqrt{0.24\times0.76}=0.50$

另一种计算区分度r_b的方法是

$$r_b=P_H-P_L$$

P_H为高分组的难度，P_L为低分组的难度。仍以一（3）题为例，算得$P_H=0.55$，$P_L=0$ （参阅附表）

所以$r_b=0.55-0=0.55$

两种算法一般会有差异，但差别比较小。

2. 区分度指标：

0.40以上的为非常优良题；

0.30—0.39为良好题，若作修改是更好；

0.20—0.29为一般题，要作修改；

0.19以下的为劣等题必须作修改，方能使用。

此试卷的八道大题共29个测题中：

区分度在0.40以上的优良题是20题，占总数74%；

区分度在0.30－0.39间的良好题是6道，占总数21%。

这两类测题占总数达95%，从区分度的要求来看，此试卷也是好或较好水平。

区分度在0.29以下的测题是一（9），四（2）及二（7）。前两道的区分度十分接近0.3，略作修改就可作为较好测题。而二（7）的区分度是－0.13，高分组的得分率反而低于低分组的得分率，是反常的，此题必须淘汰。

（六）计算信度

关于试卷的信度本书已作了介绍，此试卷的信度是采用α系数法计算。即$r_{信}=\frac{k}{k-1}(1-\sum_{i=1}^{k}s_i^2/s^2)$

其中，k是试卷的题数，k＝29；

s是试卷得分的标准差，s＝17.79；

s_i是各试题分数的标准差。以计算得：

$s_1^2=0.1204$，$s_2^2=0.1743$，$s_3^2=0.1824$，$s_4^2=0.3968$，

$s_5^2=0.4294$，$s_6^2=0.3386$，$s_7^2=0.2356$，$s_8^2=0.2158$，

$s_9^2=0.0694$，$s_{10}^2=0.2419$，$s_{11}^2=0.1971$，$s_{12}^2=0.1676$。

（以上是第一题的12小题）

$s_{13}^2=0.9984$，$s_{14}^2=0.36$，$s_{15}^2=0.9996$，$s_{16}^2=0.9639$，

$s_{17}^2=0.51$，$s_{18}^2=0.9159$，$s_{10}^2=0.9775$，

（以上是第二题的7小题）

$s_{20}^2=2.0577$，$s_{21}^2=3.2306$，（第三题的2小题）

$s_{22}^2=4.330$，$s_{23}^2=1.9448$，（第四题的2小题）

$s_{24}^2=2.740$，$s_{25}^2=5.009$，（第五题的2小题）

$s_{26}^2=7.412$，$s_{27}^2=11.286$，（第六题的2小题）

$s_{28}^2 = 8.6$（第七题），$s_{29}^2 = 1.719$（第八题）

所以 $r_信 = \frac{29}{28}\left[1 - \frac{0.1204 + \cdots + 1.719}{(17.79)^2}\right] = 0.847$。

因此，该试卷的信度已经接近优良水平。

（七）计算效度

关于试卷效度的计算，本书也作了介绍，但只是从统计的角度介绍了可以量化的准则关联效度。试卷效度还应包括内容效度与结构效度（一般只用于心理测验，教学测验不用）。

因此在分析试卷效度时，除了要确定效标来计算试卷的准则关联效度外，还需请专家或有丰富经验的教师来对试题的内容是否符合教学大纲的要求，是否复盖了教学全部重点等方面作出定性的评价——定性分析内容效度。

某市的"中考"，采用跟踪的方法对录取在高中（包括职中）的103名学生于高一上学期末举行市的统一考试，作为效标，以计算"中考"的效度。结合座谈会对试卷内容效度的定性评价—良好，作出对"中考"试卷效度的综合评价。

（八）对两类学校的教学情况的比较分析

把（A、普中），（B、职中）两类各100名考生仍按高分到低分的顺序登记到统计表上（前27人是高分组，后27人为低分组）。并分别计算（A），（B）两类学校的各100名考生的平均分、标准差、及命题的难度与区分度。

1、作整体比较：

算得此试卷对（A）类学校的平均分 $\bar{X}_A = 56.1$，标准差 $S_A = 14.2$。

此试卷对（B）类学校的平均分 $\bar{X}_b = 50.3$，标准差

$s_B=20.36$。作均数差异的显著性检验：

$$\text{因为：} Z=\frac{56.1-50.3}{\sqrt{\frac{14.02^2+20.36^2}{100}}}=2.337$$

$Z_{0.05}=1.96$。所以$p<0.05$即有显著差异。

说明这两类学校的数学成绩有明显差异。而从上分析知：试卷分数的频数线是呈正态的，难度与区分度是达良好水平，信度接近优良水平，内容效度也是良好，所以是一张比较好的试卷，因此两类学校的显著差异，不是试卷问题的缘故，而是两类学校的自身的问题，应作具体分析。

2、作各题间的比较

这个比较主要是通过它们的难度的比较（差异检验）而得。

例如由计算得此卷第一题在（A）类学校的难度 $p_A=0.78$。在（B）类学校的难度$p_B=0.42$。

因为，难度是比例数，所以采用比例数差异的显著性检验。

$$\text{统计量}\quad Z=\frac{|p_1-p_2|}{\sqrt{\bar{p}\cdot\bar{g}+\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

其中p_1p_2分别是两个要检验的比例数。

n_1n_1分别是p_1p_2所对应的样本容量数。

$$\bar{p}=\frac{n_1p_1+n_2p_2}{n_1+n_2}\quad \bar{q}=1-\bar{p}$$

若$n_1=n_2=n$时

所以 $Z=\dfrac{|p_1-p_2|}{\sqrt{\bar{p}\cdot\bar{g}\cdot\dfrac{2}{n}}}$，其中$\bar{p}=\dfrac{p_1+p_2}{2}$

对于第一题是

$$Z=\frac{|p_A-p_B|}{\sqrt{\dfrac{p_A+p_B}{n}\cdot\left(1-\dfrac{p_A+p_B}{2}\right)}}$$

$$=\frac{0.78-0.42}{\sqrt{\dfrac{0.78+0.42}{100}\cdot\left(1-\dfrac{0.78+0.42}{2}\right)}}$$

$$=\frac{0.36}{\sqrt{\dfrac{1.2}{100}\times 0.4}}=5.196$$

$z_{0.01}=2.58$，所以$p<0.01$。

即这两类学校第一题的难度是有十分明显的差异。（B）类学校的正确率仅达42%，而第一题是偏重于基本概念方面的考察，这说明（B）类学校在基本概念方面的教学是比较有问题的。

（九）写出分析报告

分析报告的内容应包括试卷的质量情况，试卷与试题应作的修改，从全市及两类学校的两个方面具体分析教学中存在的问题，提出建议。

一九八五年高中入学考试数学试题统计表

（$n=200$，$\overline{X}=532$，$S=17.79$）

姓名	考号		性别	学校	难度PH	高分组难度PH	低分组难度PL	正分度Rb
×××	150999		男	二中A				
总计	一	18	16			0.16	0.58	0.45
	二	14	12		0.41	0.23	0.57	0.36
	三	10	10		0.26	0.12	0.52	0.43
	四	12	11		0.27	0.13	0.52	0.40
	五	12	12		0.34	0.11	0.61	0.52
	六	14	14		0.53	0.18	0.86	0.68
	七	10	90		0.74	0.49	0.99	0.50
	八	10	9		0.95	0.88	1.00	0.31
	总分	100	94		0.49	0.29	0.70	0.40
一	1	1	1		0.14	0.02	0.33	0.40
	2	1	1		0.22	0.06	0.54	0.45
	3	0	0		0.76	0.45	1.00	0.61
	4	2	2		0.19	0.08	0.46	0.41
	5	2	2		0.25	0.13	0.47	0.36
	6	2	2		0.32	0.24	0.49	0.30
	7	1	1		0.38	0.07	0.70	0.61
	8	1	1		0.69	0.44	0.87	0.42
	9	1	1		0.70	0.00	0.19	0.27
	10	1	1		0.42	0.20	0.70	051

一九八五年高中入学考试数学试题统计表

（n= 200，$\overline{X}$= 532，S= 17.79）

姓名	考号		性别	学校	难度 PH	高分组难度PH	低分组难度PL	正分度 Rb
×××	150999		男	二中A				
一	1	1	1		0.27	0.04	0.61	0.68
	2	4	3		0.37	0.17	0.59	0.43
二	1	2	×		0.52	0.24	0.83	0.58
	2	2	√		0.10	0.00	0.24	0.30
	3	2	√		0.51	0.20	0.78	0.58
	4	2	√		0.60	0.35	0.76	0.43
	5	2	√		0.42	0.00	0.35	0.35
	6	2	√		0.35	0.15	0.48	0.40
	7	2	√		0.57	0.69	0.57	0.13
三	1	5	5		0.24	0.11	051	0.45
	2	5	5		0.28	0.14	0.43	0.40
四	1	6	5		0.42	0.21	0.71	0.50
	2	6	6		0.17	0.05	0.37	0.29
五	1	6	6		0.21	0.08	0.43	0.39
	2	6	6		0.48	0.14	0.80	0.65
六	1	6	6		0.43	0.16	0.88	0.65
	2	8	8		0.57	0.20	0.8	0.70
七	10		10		0.80	0.49	0.96	0.50
八	10		9		0.95	0.12	0	0.31

第二节 “SPEI”图表分析法

试卷分析方法往往得出二重性的结论。即对于同一个结果，既可能是考生方面的问题，也可能是试卷本身的问题。如某地区统考成绩偏高，其原因可能是该地区考生水平较高，也可能是考试题过易，反之，则可能是考生成绩偏低或者试题过难。但究竟是什么原因，要绝对分离这两种因素是较困难。1969年，日本应义塾大学藤田个一教授提出一种方法，在许多方面，解决了一些问题，有一定的价值。本书重点在这一方法的基础上的新结论。

一、“SPEI”图表的制作方法

首先，按题照某次考核的考生原始成绩作出原始数据表，并把每个考生的各道题目得分数按“1”、“0”填入相应的位置（对者记“1”，错者记“0”），并且求出每个考生做对的题目数和每道题目做对的考生人数。

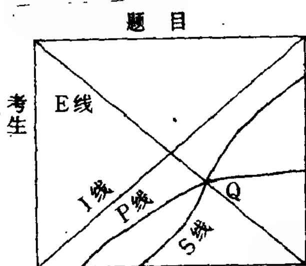
某次，根据考核成绩原始分数数据表，将考生与题目的序次按统计结果重新排列，在方格纸上制作“SP”表。纵列为考生，按成绩多次由高到低排列；横列为题目，从左到右按每道题通过人数的多少为序排列，通过人数越多的题目越靠左边，如表一所示。

再次，在“SP”表上作出“考生答对题目线，亦即绿线，简称S线。表上的阶梯形实线即为S线。作法为：先在每个考生得分横栏上相对于其答对题数的位置处通一条竖线，如名次为第6名做对10题，便在通过人数序为第10的那道题

记分位置的右侧画上一道短竖线，然后再将这些竖线（每个学生的得分线）首尾相接，即得阶梯形折线即为S线。

第四，在"SP"表上再作题答对人数线，简称P线。作法与作S线相类似，先根据每道目的答对人数n，在成绩名次为第n名考生的该题成绩的下面画一条横线（虚线），然后把这些横虚线首尾相连，即为P线。见表一所示。

当题目数与考生人数很多时，S线与P线趋近于呈两条弧线，见图一。



图一：SPEI图表简图

第五，在"SP"表的基础上作出E线和I线。从数据表的左上角到右下角画一条对角线，此条对角线称为主对角线，即平均值轨迹线，简称E线。另一条从右上角画到左下角的对角线称为I线。以S线与P线的中心交点为Q点（平均值点），Q点一般在E线上，有时也可能在接近E线处。当S线与P线有几个交点时，则以其中最接近E线的一个点作为Q点。到此，S线，P线，E线，I线和Q点已全部作出，"SPEI"表基本完成（见图一）。

表一　“SP”表

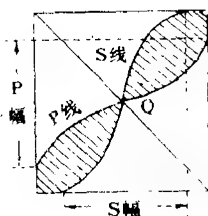
考生人数 \ 序次 \ 题目的题号		2	4	5	6	3	8	9	11	7	12	10	13	15	1	14	答对题数
	序次	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15
12	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15
4	3	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	12
9	4	1	1	1	0	1	1	1	0	1	1	1	1	0	1	0	11
11	5	1	1	1	1	1	1	1	0	1	1	1	0	0	1	0	11
5	6	1	1	1	1	1	0	0	1	1	1	0	1	0	1	0	10
10	7	1	1	1	1	1	1	0	1	0	0	0	0	1	0	1	9
6	8	1	1	0	1	1	1	1	0	1	0	0	1	1	0	0	9
13	9	1	1	1	1	1	0	1	0	0	1	1	0	1	0	0	9
7	10	1	1	1	1	0	1	0	1	1	0	0	0	0	1	1	9
14	11	1	1	0	1	0	1	1	1	0	0	0	0	0	0	0	6
2	12	1	1	1	0	0	1	0	1	0	1	0	0	0	0	0	6
8	13	1	1	0	1	0	0	1	1	0	0	0	0	0	0	0	5
1	14	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	3
答对人数		14	13	11	11	10	10	9	9	8	7	6	6	6	6	4	130 / 130

二、“SPEI”图表的基本性质

（一）、S、P线的性质及分析

如述S线实质上是一条考生成绩分布线，S线的形状如果越陡，就说明该考生成绩的差异越小，越平坦就说明考生成绩差异越大，两极分化现象严重。P线实质上是一条试题难度的分布线，P线形状陡与坦，则说明各道试题之间难度差异悬殊大与小。S、P线的形状，幅度以及两线之间所夹面积大小（即图一中阴影部份的面积大小）都是衡量分布好坏的重要质量指标。

要具体定量计算，还可以结合幅度值的推算，如图表所示，S线的幅度指的是考生成绩的分布范围，亦即表示了考生之间的成绩分布的离散程度。而P线的幅度则是表示各道试题之间难度的相对差异程度。因而，可以利用S线的幅度值大小，去比较不同集体的考生之间成绩的差异大小；利用P线的幅度值大小，去判别不同试卷的难度分布的差异情况。



图二

S线与P线之间所夹面积大小，实质是反映了考试信度的低高，亦即该次考试真实反映考生实际水平的可靠程度。S线与P线之间所夹面积越小，说明该次考试的信度越高。如果S、P两线之间所夹面积为零，即是S与P线完全重合，则说明考生都考出了自己实际水平，否则说明该次考核信度越低，没有能真实反映学生的实际水平。因此，只要根据S、P线之间所夹的面积大小，就可以推知该次考试的信度高低

了。

为了便于比较不同考核的信度低，常用下列公式定义信度r。

$$r=\frac{\text{S线与P线所夹面积}}{\text{整个“SPEI”图表的面积}}$$

例如，根据表一上的数据，采用数格可以计算得到整个“SPEI”图表的面积：14×15

S线与P线所夹面积：30

因此，$r=\frac{30}{14\times15}=\frac{1}{7}=0.143$

一般来讲，r处于0.1—0.2之间。如果r值过份大于0.2则该次考试的可靠性，稳定性就不行了。

观察表一可以看出，S线左边为“0”与S线右边为“1”的记分数目表是相同的。这两者的数目越多，表明考生越没有考出应有水平，成绩越不稳定。因此，可以用下列公式计算每个考生考查的成绩稳定系数。

$$r=1-\frac{\text{S线左边为“0”的数目（或S线右边为“1”的数目）}}{\text{试题总数}}$$

$0\leqslant r\leqslant 1$

1° 当S线左边全部为“1”（同时S线右边必然全部为“0”），r=1，表明该考生考出了自己的全部水平。

2° 当S线左边全部为“0”（同时S线右边必然全部为“1”），r=0，表明该考生根本没有考出自己的水平。

3° 一般讲，r在0.8—1.00成绩稳定。

根据表一得知，成绩相对最稳定是3号和12号考生，

$r=1-\frac{0}{15}=0$，成绩相对最不稳定是13号考生，

$r=1-\frac{0}{15}=0.867$。但他们成绩都算是稳定的。

同样，可观察到P线上方为“0”的数目与P线下方为“1”的数目总是相同的。这两者的数目越多，说明该题目越不能区分出考生水平的高低，即成绩较好的考生答不出这道题，而成绩较差的考生答对这道题的情况越多。因此，可以用下述公式计算每道题目的区分度D。

$$D=1-\frac{2P线上方为“0”的数目(或P线下方为“1”的数目)}{考生人数总和}$$

$-1<D<1$

1° 当P线上方的“0”的数目为0时，D=1，表明该题的区分度为最强。

2° 当P线上方的“0”的数目很多，与考生人数接近，D接近-1，表明该题的区分度为最低。

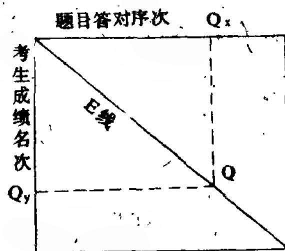
3° 一般讲，D在0.50—1.00试题区分度较好。

根据表一得知：相对来讲4题的区分度最好，$D=1-2\frac{0}{14}=1$相对来讲11题的区分度最差，$D=1-\frac{2\times4}{14}=0.429$

（二）Q点与E线的性质及分析

S线与P线的中心交点即为Q点，Q点是一个近似的平均值点。如果“SPEI”图表看作一个坐标系，那末Q点的坐标值代表了近似平均值。误差很小，在一个计分单位以内。

Q的横坐标值Q_x——近似表示考生的平均答对题数，纵坐标值Q_y——近似表示试卷中每道题平均答对人数。



图三　Q点的纵横坐标

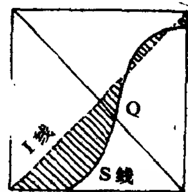
由Q点出发分别向左向上作垂线与"SPEI"表边框线相交。上面边框上的交点在边的题目难度排列序次数为Q_x；左面边框上的交点上方的考生成绩名次数为Q_y值。如图三所示。例如，根据表一上的Q点的位置，可以得到$Q_x=9$，$Q_y=8$，即考生平均答对约9题。每道题目平均答对人数为8人。

（三）I线的性质与分析

I线实质上是一条理想分辨力线。因此，S线与I线越靠近，说明试卷的分辨力越好，考生好坏差距的分辨率越高；反之，则表明试卷的区分度越差。可以用下列公式计算分辨力的指标值I。

$$I=\frac{\text{S线与I线所夹面积（即图二中阴影面积）}}{\text{整个“SPEI”面积}}$$

I值越大，分辨力越差，I值越小，分辨力越好。



图四　I线和S线所夹面积的示意图

具体计算S线与I线所夹面积时可采用数格法，即可以计算直线I线与阶梯形折线S线之间的计方格的数目。

一般讲I＞0.2时，该试卷的分辨力是不好的，不能把考生的水平差距分辨出来。如表一所示的"SP"表上添上I线后，可以计算I线与

S线之间所夹计分格子为23个，整个图表共有15×14个。因此，试卷的分辨力（即区分度）为：

$$I=\frac{23}{15\times 14}=0.109$$ 。表明该试卷的分辨力是较好的。

三、"SPEI"图表分析的作用

通过上述基本性质的讨论与分析，可以根据"SPEI"图表的分析，获得丰富的考试信息：

1、根据S线的形状，可以估计考生总体的成绩分布状况和整个平均水平。

2、根据P线的形状，可以估计各道试题的难易程度及整个试卷的难度。

3、根据S、P的幅度值大小，可以大致了解考生成绩及试题难度差异的离散程度。

4、根据考生人数与每道题目在P线上方的"0"数目，可以估计每道题目区分考生水平高低的作用大小。

5、根据S线P线之间所夹的面积及整个图表的面积可以估计试卷的信度高低。

6、根据题目总数与每个考生在S线左边的"0"数目，可以了解哪些考生没有考出其实际水平。

7、根据S线与I线之间所夹的面积及整个图表的面积可以估计试卷的辨别考生的差异的作用大小。

注意一点，"SPEI"图表所提供的信息，虽然丰富，但还是评价试卷质量的常用指标。而且要特别提醒的是它局限于考试的试题必须每题占分相等，一般需采用"0—1"记分法。对于不同占分的题目如何使用，权重如何安排，有待

于进一步研究与探索。另一个问题是当考生和题目数量增多时，制表就困难，这需要电子计算机的帮助。对于规模较大的考试还是统计分析方法占优。

第九章　考试与智力测验（简介）

考试不仅是教学工作的基本环节，“是检查学习和教学效果的一种重要方法。”而且对“教”和“学”的要求、方式和方法还具有指导作用。这种指导作用是潜在的。考试本身，既蕴育着对教师教学的要求，也蕴育着对学生学习的要求。考试实质上是一根无形的“指挥棒”它具有智力训练，引导应试者创造地学习，开发智力等意义。换句话讲，考试是一种高强度的智能训练。它可以训练应试者的审题能力、运算能力，验证能力、表达能力；它可以训练应试者的沉着、冷静、细致、严密、敏捷、有序、灵活的科学思维素养。

第一节　智力的概述

一、智力的定义

什么是智力？一百多年来，国内外心理学家、社会学家，哲学家等都在研究，至今没有一个满意的定论。现在更是众说纷纭，莫衷一是。

有不少人从生物学、心理学、行为科学等多方面去给智力下定义。但是任何从单方面去定义智力，都难以为人们所公认。生物学方面认为智力是“中枢神经系统的功能”，心理学方面看，智力是“进行抽象思维的能力”。有人又定义

为：“获得能力的能力”；“天生的一般的能力”。正如《大英百科全书》所指出的，这些定义并没有被人们所普遍接受。从行为科学方面看，给智力以“操作定义”，认为“智力就是智力测验所测量的那种东西”。法国心理学家比纳认为：智力是“善于判断，善于理解和善于推理的能力”。瑞士心理学家皮亚杰在《智力心理学》中提出“智力实质上是一个活跃的和积极的操作系统，它是最富于推动力的心理适应作用，也就是说，是主体与环境之间进行交流所不可缺少的工具。”“智力活动实质上就是在于按照某些确定的方式‘组合’各种操作”。[①]《辞海》对“智力”是这样下定义的：智力“通常叫智慧，指人认识客观事物并运用知识解决实际问题的能力。集中表现在反映客观事物深刻、正确、完全的程度上和应用知识解决实际问题的速度和质量上，往往通过观察、记忆、想象、思考、判断等表现出来。它是在掌握人类知识经验和从事实践活动中发展的，但是不等于知识和实践。它是先天素质，社会历史遗产和教育的影响以及个人努力三方面因素相互作用的产物。”这一定义基本上明确了“智力”的外延和内涵。从教育的观点来看，智力是什么呢？智力“是在遗传素质的基础上，人认识和改造世界的实践活动中的心理特征和各种能力综合表现”。[②]总起来讲：智力被视为一种潜在能力，这种潜在能力只能是遗传，后天发展和成长的一种[illegible]。因此，作为一种潜在能力，它也会由于疾病或受到刺[illegible]化，这同一个人的其它身体特

①杨清，《心理学概论》，吉林人民出版社，1981，第592－596页。

②查有梁，《控制论、信息论、系统论与教育科学》71页，四川省社会科学院出版社出版。

征是一样的。另一方面，智力成长为潜在能力的过程，可因环境的压力而受阻。也可因适当的刺激而加速，但是增长速率的大小或停止进一步发展的时间的早或晚，均不会改变潜在能力本身。潜力的实质在于高效地发挥个体在认识和实践活动中的自觉能动性。

二、智力的结构

智力的结构理论反映着智力的本质和定义，只有解决了智力结构理论，智力测验才有理论依据，也只有解决智力结构问题，才能对智力和其它特征（如学习效果、生物影响，才能的结构以及智力的年龄变化等）进行深刻的研究。由于智力的定义本身没有彻底解决，因此，智力结构问题也一直有争论。归纳起来，围绕智力有多少独立可分的因素，以及这些因素又是以怎样的关系展开的。其理论，可以分三部分，下面简述之：

（一）比较原始的观点

古代，把智力、能力、知识、技能甚至经验混为一体，不加区分。但是对智力及其结构有其观点。一种认为：智力是稳定的、不变的。而智力的减退是由于生理变化引起的（如大脑损伤、营养不良和疾病等），智力是脑中早已存在的东西，是天生的，是由遗传决定的，与文化没有关系。柏拉图就认为智力是天赋的。中国古代的孔子、孟子也是持该观点的。另一种对立观点是认为智力就是后天学习的结果，是由文化和个体经验决定的，知识起着决定作用。

（二）智力结构两因素理论

因素分析是一种统计技术，它通过对测验作业的相关因

素、相关系数去发现智力功能，然后再分离出这种功能的基础能力。通过许多测验所得的相关系数，就可以辨别出其中的共同因素，然后，按照这几个相关测验所需要的能力，来给这些因素以名称。如视觉记忆能力。数的运算能力等。因素的分析可以说明一个人有多少能力。由于有了因素分析技术，就产生了几种智力结构理论。

英国心理学家斯皮尔曼（Spearman）首次用因素分析法来解释智力结构。他认为智力主要由两个因素构成，一是一般因素（或称心理能量mental energy）用g表示，它渗入到所有的智力活动中，人人皆有之，可人人又不相同。另一个是特殊因素（或具体因素specifie factor）用f表示，这一因素数量大，与特定的任务高度相关。斯皮尔曼认为g代表一般的心理能量，是天赋的一部分，他把s比作机器或引擎，g是能量，机器则由能量来使其运转。其中s受教育和训练的影响很大，而g是先天的，非教育的因素。他的三个著名认知规律经验顿悟（apprehension of enperience），关系推演（eduction of rection）和相关推演（eduction of conclutions）的理论的根源就来自于此。

斯皮尔曼智力结构两因素的拥护者卡特尔（Gattell）通过测验和因素分析研究，认为斯皮尔曼因素不是一成不变，也不能归结为天赋，提出g由两部分组成，一部分是不定形智力（fluid interligence）用gf表示。另一部分是已定形智力（Crgrtalliecl interligence）用gc表示。他称gf和gc是斯皮尔曼g因素的双生子，是其它因素中的因素。

他认为gf和gc并不是一个天生的智力，而是一个获得的智力。尽管gf和gc更具有天生的性质。但它也取决于环境的

影响，尤其是脑和其它神经系统受到的有效影响。gc是后天获得的，是早期gf活动的结果，是学习的结果。它可以在好的教育条件下更好的产生，因此，gc有知识的成分，最起码也有经验因素参与。而gf实质上对复杂关系的推演，具有纯净能力的功能，它包含于所有对复杂关系的认知的操作中，而这种操作不能借助于记忆，已定形能力和判断技能来完成。从上述观点看，gc是gf的载体，所以卡特尔称他的理论是投资理论，原因是已定形能力成了一不定形能力的"信托人"而gf就是"投资者"。

他认为由于各种因素的作用不同，那么所处的层次也就不同，一定的因素稳定在某一水平位置上活动，任何行为都包含有一定的能力，也就是说你作了什么，那么你就有相对应的某种能力。因此，gf和gc的具体结构是比较复杂的。

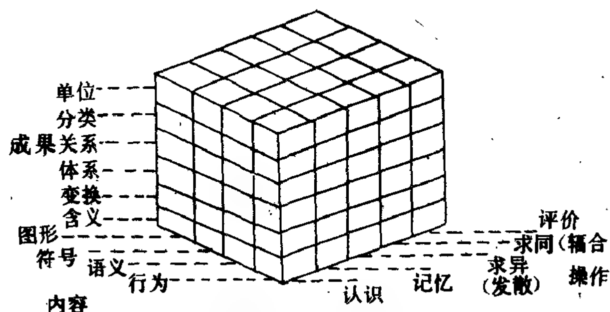
卡特尔的理论是斯皮尔曼理论的发展，他从gf和gc两个方面阐述了g因素的实质，g因素和s因素的关系，能力和行为的关系，他的g因素渗透于各种能力之中的思想，行为是能力载体的思想、反对g因素天赋的思想是可取的。但是，他想寻找零载体的g因素，却违反了个性与共性相互依存的原理。

（三）智力结构的群因素理论

美国芝加哥大学的色斯登（L. L. Thurstone）认为智力结构不应是两因素理论，而智力应是在四种主要水平上的试误，外在试误（最低智力水平）知觉试误、观念试误、概念试误（最高的智力水平）。例如一个人沿着大马路散步，在随便行走中，知觉试误智力就够用了。但当一看到有汽车迎面而来时，观念的试误智力就要起作用了。因为观念的智力

有提前试误的作用，他认为概念试误智力是以大略的，组织松散的，不完全活动（即简缩的）方式进行的。而色斯登所测验和分析的智力正是这一水平的智力，通过他自创的因素分析，他发现，一些测验并没有g因素，而只有一些元素性因素，这就使他得出结论，智力是由许多左右的因素构成的，他称之为原始心理能力，其中最主要的有语言能力（V），数能力（N），空间关系能力（S），言语流畅性（W），记忆（M）和推理（R）等。根据这些因素他编制了"芝加哥原始心理能力测验"但是与他的愿望相反，并没有发现这些能力是独立的，而每一种能力实际上与其它的每一种能力都有正相关。于是他就解释说，这是第二位因素在起作用。实际上这些结论说明，可以从原始能力中找出更一般的因素。因此，他的理论为更广泛分析智力结构开辟了新途径，他的工作促进了后来的群因素分析研究和理论的发展。

美国心理学家吉尔福特（Gwilford）对来自各种因素分析的模型都感到不满意，经广泛的测验研究，提出了智力三维结构模型。使群因素理论得到进一步发展，他认为，智力的结构应从内容、操作和成果三个角去考虑。如图所示：



吉尔福特提出“内容”的项目有四种，“操作”的项目有五种，“成果”的项目有六种。这样在理论上就可能有5×4×6=120种可能的智力因素。而每一种智力因素就是一种独特的能力，都由心理操作，内容、产物构成，即模型中的每一个小立方体。

现在很多人认为这一理论是对智力结构认识的一个深入，它扩大了对心理能力分析的范围，对编制智力测验，对教学实践都有一定的用途。

波特（Burt）和沃南（Vernen）这两位心理学家对斯皮尔曼的g因素进行了深入具体的描述。他们认为对智力结构中不断发现的新因素，进行解释的最好方法是使用树状层模，按逻辑的，交互相关的关系从g因素→群因素→具体因素（S）来摆放这些因素。

波特的模型（1949）：波特的智力概念，指的是整个人类的心理能力，他又把心理能力区别为能力——本质是智力，采用斯皮曼的g代表，和实践——本质上是行为，用P代表。后者包括心——动能力（即心理对活动的调节）机械和定向能力。按树状顺序排列这些因素，认为每一高水平的因素直接来自两个次一水平的因素，叫做两分枝。第一个两分枝的头是人类心理能力，两分枝端所处的水平叫关系水平，即g因素和p因素水平。g因素和P因素，又各自分两枝，分枝端点所处的水平叫联合水平。依此类推，在g方面的第三水平是知觉水平，最后是感觉水平。很明显波特的模型最终是用具体因素来解释g因素的。

沃南的模型（1951），他也认为智力是由g作为头的树状层模型，g往下分成两个因素群，一个是言语——教育因

素群（Verbal－educational）简称V.ed，这种言语——教育的复合能力涉及到类似那种学校里处理作业的能力，以及类似成就测验的书面作业能力等。另一个是动觉——运动因素群，简称K.m，后者相当于波特的实践因素。这两个因素群又分成下层更小的因素，如V.ed分为语言因素，教育因素等，K.m分成实践因素，空间因素和机械因素，生理因素等，如此下去一直分到具体因素S。沃南认为智力与一个人一生中所积累起来的图式，复杂性和灵活性的概括化水平相一致，它的获得要受先天能力的限制。高一层次图式的出现依赖于具体知觉图式的获得，所以图式的水平越高（概括化程度越高）g因素所包含的就越多。沃南认为图式的丰富性和灵活性是随环境而变化，随环境而确定。

波特和沃南的理论既发展了两因素理论，又有群因素理论的成分，把g因素和S因素联系起来考虑，这是比较合理的。

总述起来，智力结构理论认为，智力一部分是特殊智力或专门智力，一部分是一般的智力；一般智力不是铁板一块，它要么是由若干个，要么是由若干群，要么是由若干层更低一级的因素构成；思维能力，思维操作的特点作为智力的核心。

第二节　考试与智力发展的辩证关系

“考试期间复习巩固知识是天经地义的，何谈‘提高’，‘加深’，至于发展智力哪能依靠考核呢？”这样谈论不只是非教育工作者，而连有的教师也有如此之看法。这里我们

可以大胆地讲，如果通过考试并不能促使应试者所学得的知识加深和巩固，不能对学生的智力发展有所促进，那么考核的目的就不能说已经完全达到，古人云“温故而知新”。复习和巩固决不意味着只是在原地踏步。事实上，学习和掌握知识的过程本身就包含着深化和发展，何况考核并不仅仅是一种简单的“温故”。有人说：“质量不是考出来的，而是学出来的，教出来的”。这话诚然不错，但却不能不看到，学生应如何学，教师应如何教，常可从考试和测验中得到启示，应试者的独立思考和主动探索精神也会在考核的过程中受到培养和锻炼。考试对应试者思想品质，学习目的和态度的检验更是不可忽视的。那种认为“考试只能说明学习”的见解是狭隘的。我们认为：考试对应试者的智力发展有着很大的关系，在某种程度上决定着应试者的智力发展的方向，范围、结构、层次、规格和类型等，直接影响着人才的培养与发现。

一、考试与智力发展的内在联系

现实生活表明，考试的目的性决定了应试者的目的性，考试内容与标准决定了应试者的知识结构与智力规格。从考试的职能看，它既是检验教学效果，调查和改进教学方法的有力措施，又是鉴定和选拔人材的比较可行的手段。学校把考试视为选拔新生和鉴定学生的学习成绩的好坏的主要手段之一。国家劳动人事部门把考试作为考核干部职工的重要方法。考试在选拔人才中所起的特殊作用，往往给人们成一种心理趋势，好象整个学习的目的就是为了考试。从而看出考试对应试者（特别是在校学生）智力的发展产生了一定的向

心力。由于考试具有竞争性，应试者为了不被淘汰，常常是自觉或不自觉地按考试的标准与要求来调整自己的知识和能力结构。造成应试者的智力在哪方面受到刺激，就必然引起哪方面的急剧发展。一个人要想成为社会所需要的合格人才，就必须接受国家的挑选。目前，考试就是一种衡量人才的“客观”准绳，为此，人们就不得不按照考试指挥棒来塑造自己，形成与考试要求的标准相符合的智力结构。这样，就使考试与智力的发展产生了一种内在的直接联系的影响。考试给应试者无形中提出了智力发展的方向、范围、结构、层次、规格、类型等一系列要求与规范、制约、影响应试者的智力结构。

二、考试与智力发展的关系

考试有什么标准与要求，应试者就会朝什么方向努力，智力就会朝其方向发展。然而，考试的标准与要求又由什么决定的呢？它是有一定的客观依据的。首先，考试的标准必须符合生产力和科学技术发展对教育所要培养的人才提出的客观要求。各个时期生产力和科学技术发展的水平决定着该时期对人才的要求的标准。从而也为该时期考试提供了一定的客观依据。

其次，考试的标准必须适合人的智能发展的自然条件（生理因素），从生理学、心理学的角度看，人的智力发展有它自身的规律，如果考试标准适合人的智力发展规律，就能引导他们形成最佳智能结构，反之就会使他们的智力发展受到压抑、制约或畸形发展。科学研究表明，人的大脑可分为四个功能区，即直觉功能区、记忆功能区、判断功能区和

想象功能区。从这四个功能又引出，人的智力概括为注意力、观察力、记忆力、思维力、想象力、创造力等，心理学研究表明，人的智力的各种因素是相互联系、相互制约、相互影响的。智力活动是各种智力因素共同发挥作用与和谐统一的过程，其中任何一种功能因素的过分扩张与发展，都会造成大脑功能的不协调运动和畸形发展。一种功能的恶性膨胀，必然会引起另一种功能的压抑与衰萎。因此，大脑功能的运用正如人体的正常发育一样，必须保持一定的平衡与协调，考试在智力发展上则起着调节、诱导和规范的杠杆作用。要使智力得到全面的发展，就必须改变片面地刺激和强调某种功能，而应全面地考试。

考试与智力发展的关系主要表现在三方面：其一，它关系着智力发展的方向和范围，考试既可引导智力朝纵向发展，又可朝横向发展；既可单一发展，又可综合发展。从人类的智力发展史看，智力经历了由综合向单一，又由单一向综合的否定之否定过程。原始社会生产力极其低下，几乎谈不到科学技术，人的智力结构是自然趋向综合发展。随着生产力和科学技术的发展，社会分工越来越细密，科学分科也越来越具体，客观上要求人的智能由综合向单一，由横向向纵深发展。现代科学的发展和电子计算机的普遍应用，生产的社会化与科技的整体化，又要求人的智能由纵向向横向，由单一向综合发展。这从客观上确定了考试的新标准，使之在选拔人才的过程中重新调整人们的智能发展方向。目前，出现了许多边缘科学，有些高等院校根据这种趋向确定文理科课程纵横交错，互相渗透，就是为了改变学生的智能发展方向的单一性。

其二，它关系着智力发展的结构与层次。自然界是一个多层次，多序列的网络结构，科学发展的不同阶段，要求人们的智能结构也是不尽相同的。最佳智力结构应该是能够适应各个时期生产和科学技术发展的趋势和特点。现代科学发展的特点出现了许多边缘科学，自然科学与社会科学在某些领域出现了立体交叉，相互融合的趋势。历史上旧科举制度之所以培养出脱离实际死啃书本的书呆子，根本原因就是偏重知识的机械记忆，引起智能结构畸形发展。现代科学要求我们全面地培养学生各种能力因素的综合运用，以建立最佳智力结构，提高智力素质，克服片面发展，使学生智力发展趋向整体化，既注意专门训练，又注意综合运用，既注意平面发展，又注意立体交叉，这样互相渗透，互相补充，有很高的应变能力，以便从这一领域迁移到另一领域。

其三，它关系着智力发展的规格与类型，尽管科学的发展日益趋向整体化，但现代社会存在着分工与协作，科学的相互渗透仍然是建立在科学的分科基础之上的。因此社会不仅需要通才，也需要专才。科学的整体化与科学的细密分科是一个辩证的统一。这就既要注意到智力发展的整体化，又要注意到专一化，各个学科和分支对人才的要求不同，可以根据各自的特点确定考试的范围和标准，有些学科需要全能型，有些需要专能型，有些需要艺术型，有些则需要思维型等等。应试者可根据个人的基本条件和客观标准来调整自己的智力规格与类型。

总之，考试可以促进智力发展，也可以抑制、束缚、压抑甚至摧残智力的发展。当然，智力的发展并不是必须依赖考试的刺激和引导才能完成。许多自学成才的人，他们一方

面适应了社会生产力和科学技术发展的客观需要，一方面又注意培养和锻炼智能的发展，充分发挥个人的聪明才智，因而取得成功。

第三节　智力测验

智力测验主要是指心理学用以测量人的智力水平的一种方法。在英国心理学家高尔顿（Francis Galton 1822——1911）创始的心理测验的基础上，于1905年由法国心理学家比纳和西蒙（Theodose Simon 1873——1961）用语言，文字或图画，物品等形式，编制出一套"量表"。测验时要求受试者用文字或动作解答，然后依照公式，求出受试者的"智龄"和智商（或用其他方法计算成绩）从而确定其智力的高低。这个方法是根据当时法国教育部的要求，用来检查小学生留级的原因的。那么，现在智力测验又有什么用呢？研究它对考试有什么作用呢？

一、考试与测验

严格地说来，考试和测验具有不同的内涵。这在现代教育的文献中是很清楚的。测验比考试具有更为广泛的内容。它不仅包括了学业成绩和能力测验，还包括对智力，兴趣等方面的测验，甚至也不单单用于教育领域，应用于教学过程的考试和成绩测验，在试题内容和测量方法上也存在着明显的差别。然而，作为对表现考试的方法手段上，二者又有某些共同之处。所以经常习惯地把"正式"的成绩考核（如学年考、毕业考、升学考）称为考试，而把"非正式"的考核

（如平时考，模底考等）称为测验。国外也有类似的看法，认为“测验”评价是非正式的，它的结果不能作为正式的证明资料。

二、智力测验的兴趣

最初出现智力测验是在1869年，是英国心理学家高尔顿创始的心理测验。当时的心理测验，是用以测量人在智力水平，心理特征方面的个别差异的方法，用作鉴别学生优劣、测查犯罪原因，挑选职工和士兵等的工具。方式主要是使用实物（或器械）和文字（或图形）。1890年，高尔顿根据这些测验发表了《心理测验与测量》，创造了“心理测验”这个术语。心理测验的结果，通常用测验量表加以衡量，用统计方法加以处理，并用数字或图表等加以表明。心理测验种类很多，除了智力测验外，还有品格测验，能力测验，成绩测验等等。

十九世纪后期法国实验心理学和智力测验的创始人A·比纳（Alfred Binet 1857年生于尼斯（Nicel），1911年卒于巴黎，和他的同事V·亨利一起建立了法国第一所心理实验室，他和他的合作者花了许多的心血。对智力测验的方法进行了具有独创的和各种途径的探索，其中包括头颅、面部、手形的测量以及对笔迹的分析，通过他们艰巨又严峻的探索，对测量复杂的智力结构提供了宝贵线索。1895年在比纳倡导下，法国出版了第一种心理学杂志，即《心理学年报》。

1904年，巴黎教育当局委托比纳，承担编制一套测验，用于鉴定智力缺陷的小学生，让他们能够进入教授特殊课程的学校，为此，比纳和T·西蒙（T·Simon）合作，于1905

年创立了第一个《比纳——西蒙智力量表》，即著名的1905型量表，量表由30个难度不同的试题组成。其难度的划分是通过对50名智力正常，3－11岁的儿童以及一些智力低下的儿童和成年人的测验后决定的。这种测验的覆盖面相当广，特别着重对判断、理解、推理的能力的测验，因为这些因素被比纳视为智力组成的基本要素，测验中尽管有测试感觉的部分，但测试表达能力的部分占有相当大的比例，由于1905型量表在当时是作为一个不成熟的试验工具，因此，对各年龄阶段的儿童没有形成精确的达到总分目标的要求。

1908年比纳又发表了这个量表的修订本，即第二量表，或称1908型量表，这个修订本不但增加了试题，删去了前一量表中的一些不足的测验，而且使试题的难度随年龄的递增而上升，即所有试题是在对300名智力正常3－13岁的儿童进行实验的基础上，按年龄水平分等级。所以，3岁水平的测验是被80％－90％智力正常的3岁儿童通过的测验，4岁水平的测验是被80％－90％智力正常的4岁儿童通过的测验，如此类推直至13岁，量表的应用年龄是3岁－13岁，这样，被试儿童的得分可和智力正常的同龄儿童的得分相比较，该儿童的智力情况就可通过智力水平来表达。因此，1908型量表是第一个年龄量表。

1911年比纳发表了这个量表的第二次修订本。这次修改只作了微小的改进，增加了一些特殊的测验项目，主要是扩大了其应用范围，使该量表容量增大。比纳在这次修订本出版之前已经逝世。

尔后不久，比纳——西蒙的智力量表引起了全世界心理学家的极大兴趣，当时出现了各种文字的翻译及修订本。其

中最著名的修订本是在斯坦福大学教育心理学教授L·特曼（L·Termen 1877—1956）的指导下完成的，即斯坦福—比纳智力测验（1916型）也就是在这套测验中，第一次使用了智商（IQ），即智力年龄和实际年龄间的比率（CA）。

在比纳及其智力测验的影响下，随着心理测验技术的不断复杂化，测验也越来越精确，可靠和多样化方向发展，到本世纪三十年代，智力测验得到很大发展，瑟斯顿，吉尔福特，韦克斯勒（D·Wechsles）等人都作出了卓越的贡献。国外目前通行的智力测验，已开展起来了，智力测验对教育应用方面也有新的进展。美国心理学家布鲁姆强调考试的重要性时，主张搞"探索性考试"。美国在测量天才儿童时，采用测验一般智力和特殊才能、艺术才能等等方法，并把一些"创造性思维的测验"与传统的"智力测验"同时并用。近几年来，还有人根据某些学生"智力教育"低而实际工作能力并不差的情况，主张在"智力测验"之外，搞"兴趣测验"、"人格测验"、"能力测验"等等[①]美国大学招生，往往有两种考试，一种叫成绩考试，一种叫能力考试，能力考试也与学科知识有关，不过更浅些，更活性，能考出学生能力的高低，美国有许多大学很重视能力考试，把考试重点放在考察能力上，哈佛大学在康南特任校长时就只搞能力考试[②]英国中学也十分重视对考试工作的研究，各学区考试委员会都与一些大学联系，各自制订各科的考试大纲，采取不同的考试方法，1980年英格兰一个考试委员会的物理考试共分为四次进行：第一试，考查实验能力，时间为一个半小时，

①参见《外国教育史资料》，华东师大教育系，1973年版第200～201页

②参阅，刘佛年：《全面发展和教育改革》（教育研究）1980年第5期）

成绩占15%，第二试，考查基础知识的掌握情况，要求速度、准确，时间是一个半小时，成绩占30%，第三试是考查知识的综合运用能力，有些题目的要求难度较大，时间三小时，成绩占35%，第四试，主要考查能力的强弱，时间二小时，成绩占20%。[1]一九七九年日本开始实行新的大学招生考试制度，改变了以往一次成绩决定成败的作法。高考分两个阶段进行。第一次考高中的基础知识，第二次主要考查考生对所学专业的适应性，包括思维能力和综合能力等，各大学根据两次考试的成绩和中学提供的调查表，统一研究是否录取。[2]

如何对待智力测验，评价智力测验，使用智力测验。国际心理学界不仅看法很不统一，而且各地区性的做法也不尽相同。就拿盛行测验的美国来讲，就有许多州，立法禁止智力测验，加利福尼亚州政府1975年制定不准施行智力测验的法律条例就是一个例子。共和国成立之后，由于受苏联的影响，前二十五年心理学一直是禁区。今天，智力开发成为我国资源开发的重要课题，测验自然也开始"抬头"，由比纳所开创的智力测验自然地是考试（包括升学考试）的一种重要的辅助手段。

三、智力的分配

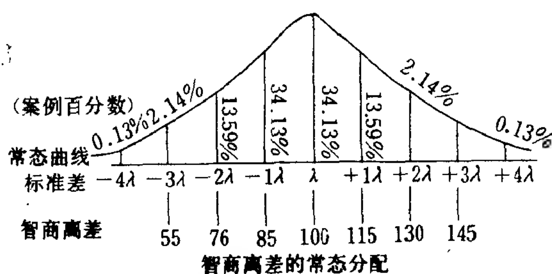
测量出来的智力也象测量出来的个人的其他特征一样，其测验结果有按照"常态曲线"分配的趋向。就是说，平均智商大约是100，得分靠近这个数字的人数最多，而得分与100的差距越大，则人数就越少。

随着所用的测验不同，则总人数的实际智商分配就多少

① 徐惠，《英国提高中学教学质量的措施》，《外国教育》1981年第十期。

② 参阅，全世伯，《日本改革大学招生制度》、《光明日报》1981年4月8日。

有所不同。自然智商分配的变化是由不同的智力测量方法造成的。对于某一智商，必须在明白了它是以何种方式获得以后，才可进行解释。假若一个儿童的智力分数仅记为一种智商分数，而不知道这种分数是在什么时候以及如何获得的，那它就可能没有多大的实际意义。我们举一个智商是90的儿童为例。假若这个智商是根据奥提斯智力快速记分测验乙种量表得出来的，则公认它比用传统的方法获得智商更靠近100，另一方面，假若这是一个在视力或阅读方面有某些困难的儿童，则这个分数可能低于他的实际智力水平，因为奥提斯测验比大多数常用的测验更多地依靠一个人的阅读能力。假若这个分数是从斯坦福——比纳测验得出的智商，则这个人大概超过一般人的百分之二十八或二十九。所有测验分数的意义都要看它们是依靠什么测验得出来的以及当时占优势的许多其它情况而定。



智商离差的常态分配

智商离差的常态分配

四、創造能力的测验

能力，通常指完成一定活动的本领。包括完成一定活动

所必需的心理特征。比如，从事音乐活动既须掌握歌唱、演奏等具体活动方式，又须形成曲调感、节奏感，音乐听觉表象等心理特征。各种活动所需的心理特征在各人身上的发展程度和结合方式是不同的。因而能力特征也是人各不同的。能力是在人的生理素质的基础上，通过教育和培养，并在实践活动中吸取了人民群众的智慧和经验而形成和发展起来的。创造能力就是指完成创造活动的本领。独创性，首创性，新奇性是它的基本要素。首创性是它的基础。研究创造能力常有三种途径，把重点分别放在创造的过程上，放在创造者上，放在创造产品上。

很多人认为创造能力是一种智力的表现。但据创造能力研究结果表明。创造能力与智力测验的结果相关度并不太高。一般而言，创造能力高的学生智力也颇高。而智力高的则创造能力未必也高。智力高低可算是创造能力高低的必要条件。但并不是充分条件。智力的问题需要依据一定的逻辑规律和运思策略去将资料组合，推断和解答；创造能力问题除了需要掌握一般逻辑和策略外，还需要想象和巧思的能力、将资料作创新性独特性组合和解答。智力问题需要聚向思维。创造力问题不仅有聚向思维，而且还有发散性思维。

创造能力测验大多要求受试者提供创制新颖的答案。吉尔福特及其同事在创造能力测验上做了大量的工作。1960年他们就已经设计出一套评价创造力的精细工作表。早在1950年，吉尔福特在美国心理学会所做的主席致词中，指出“智力的结构是复杂的，需要多样化的测量方法。他曾假定，创造能力中包含的那些思维能力就是他定义为“发散的成品和变换”（divergent Productions and transformations）的

那些能力，与他曾指出的并当时流行的智力测验所测量出来的“辐合机能”（Convergent functions）恰好相反的。后来吉尔福特一直坚持他的辐合思维与发散思维的对立观点。1962年，盖策尔斯和杰克逊在研究中应用了五种不同的方法来测量创造能力。这些思想方法有些是沿袭吉尔福德的，一些是他们自己创造发明的。

第一种测量方法是词联想测验。其中要求应试者对一个象“螺钉”或“袋子”之类的十分普通的词，尽可能地下多的定义。测验分数决定于定义的绝对数目和这些定义可分为几类的数目。

第二种方法是物品用途测验，这个测验同吉尔福特的认识能力的研究中所用的测验相似。被试者对一个象砖块或牙签之类的普通物品，尽量地说出其用途。根据所说的用途数目和首创性这两个方面来评分。“砖块可做床炉”是一个比“砖块可建筑房屋”更有独特性的回答。

第三种方法是从以前已有的一套测验中借用来的隐蔽图形测验。在这个测验中，给被试者看看一张上面有简单的几何图形的卡片。然后要求他把那个以更复杂的形态或花样而隐蔽着的图形找出来。

第四种方法是寓言测验，在这个测验中，给被试者呈现几个短寓言，但却缺少最后一行，要求被试者对每个寓言作出三个不同的结尾：一个“道德的”，一个“诙谐的”，一个“悲伤的”，根据结尾的数目，恰当性和独创性来评分。

第五种叫做组成问题。在这个情境中，给被试者呈现几篇复杂的短文。每篇短文包含一些数字说明，要求被试者根据已知的资料尽量组成许多数学问题。根据问题的数目、恰

当性、复杂性和独创性来评分。

1966年，托兰斯（E·P·Torrance）发明的创造性思维测验，相当广泛地用作测量创造力的工具。它包括这样一些作业：问与猜测验，产品改进测验，非常用途测验以及合理设想测验。这些测验提出的问题都强调不寻常的或聪明的想法。这套测验还包括拼图测验、不完全图形测验和平行线测验。托兰斯测验的四个评分标准如下：流利——中肯反应的数目；灵活——由这一种意义转到另一种意义的数目；独创性——反应的罕有性；精密——反应的详细和特殊性。所有这些作业都要求提出各种不同的解答，多种可能性以及在理论上包含在创造性行为之内的某种思维类型。

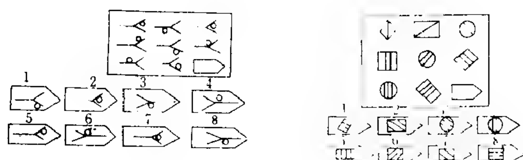
第四节　智力测验主要流派

早期心理学家认为：智力测验主要是考查应试者的理解、推理、判断能力，于是，理解、推理、判断就成为智能的要素。后期的心理学家对智力的理论，就有不同的注释了。归纳起来，主要有以下的学派。

一、心理测量学理论

心理测量学理论，主要是利用心理测量的量化方法来分析智力的特殊。它企图从智力因素结构中来了解智力的性质，早期智力测验的基本思想是探求被测者的智力表现是根植于几种智力因素。当时认为智力只是个别功能。如感知、记忆、联想、思考、想象等的综合表现。各个功能聚并而成智力机体。强的智力机体中，其各个功能亦强，反映在智力测验项

目中的各项表现就较佳。反之，则表现较差。智力表现的高低、优劣取决于智力机体的强弱。因此，智力就成为带有普遍性的单元结构。早期智力测验的研究表明，除各个智力测量项目中各自都有各自的某些独特能力外，各个智力测验项目在测验时表示出一个渗透各个项目的变项。反映智力机体的普通的智力因素。智力机体在某一种智力测验中表现较强，则智力机体的各项也应有较佳的表现。若一项弱则其它亦有较弱的表项，雷氏矩阵智力测验就是一种普通的智力测验。



智力发展是多样的，内容取材也是广泛的，不少心理学者利用“因素分析法”研究智能基因。对学生各式各样的智力测验。然后根据学生的表现进行因素分析。一般研究结果表明，每一智力测验后都能找出七八种较为独立的智能因素。同时，各个智力测验项目中的表现并不是一致的。如象数理符号等直观性强，语文理解方面则弱，而语文推理优越，亦未必在定向机械性推理方面表现超群，智力不是一普遍性的单元结构。而是一多元结构，蕴含一些性质不同，功能独特的智力因素。下列就是智能因素的智力测验。称“能向测验”。

语文推理

小安、大治：____：____　路途：崎岖：____：____

1、小心，大意	1、手工，巧拙
2、小贼　偷窃	2、沙丘　山坡
3、善良　凶恶	3、航海　蜿蜒
4、骚乱　暴动	4、命运　坎坷
5、灾难　拯救	5、运气　顺逆

由于因素分析法，对各类测验所得成绩均可为依据。亦可产生单元和多元法结构模型。有人认为某些智能因素比其它的是更基本层次的结构。亦有认为各个因素同等重要的平行结构。可见心理测量研究带来了相异的结论。无论单元、多元、层递或平行结构，各指出其论点。但是，各模式之间只是各有偏重，无实质上的差异。单元结构并未排除因素的存在。而多元结构亦未排斥智力的功能。心理测量学理论着重智力结构和智能因素的剖析。它提供了不仅是人类智能分门别类的有效工具。还是各式测验结构效度验证时不可缺少的工具——因素分析法。

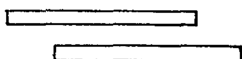
二、皮亚杰理论

皮亚杰理论是利用发展心理学的成果来研讨智力的内容和性质。他指出：不同年龄阶段的人之间智力不同是由于他们利用不同的逻辑思维形式去解说事物的变化，因而对事物规律有不同的掌握程度。各阶段的逻辑运思形式不同就影响认知结构不同，继而影响智力表现之不同。事物内在规律支配事物的变化，而掌握这规律的逻辑运思就成为智力的关键性内涵。例如，初中生若凭观察则知道了事物在水中浮沉和

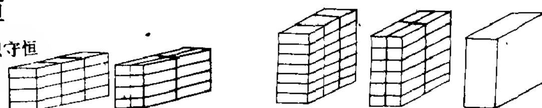
事物的大小重量有关，这只是从掌握事物的外征去运思。但若了解事物的密度才是决定事物的浮沉这一个内在规律，才算懂得逻辑运思策略，才能加强认知结构的内容。同理，体积，重量，甚至意义的守恒。例如语文中的借喻、寄意就是将意义的转移而产生守恒作用。这是语文运用的内在规律，须要语文逻辑运思形式方能掌握。因此，根据这个理论，智力的高低就由形式逻辑运思和成熟与否来决定。而智力的具体表现就由事物规律，守恒观念，类别推理等来反映。以下就是皮亚杰理论智力测验中的一些例子。

长度守恒

两根等长的棒子平行但不并放着，让被试者注意棒子之间的参差判断长短，以研究其长度可逆性的水平。



容积守恒



皮亚杰理论强调对事物内在结构和变化规律的认知，可增进逻辑的策略，每掌握一规律，即可长一智略。不同阶段的儿童掌握了不同数目和不同性质的智略，因而有不同程度的智力表现。换言之，对事物规律的掌握就成为反映智力的要素。认识问题，是一个复杂的问题，每一认识主体都处于复杂的社会联系之中。认识的产生和发展不可能不受到社会联系的制约。因此，对认识过程的研究只有放在社会联系之中并进而考察认识的发生，才能得出较为符合实际的结论。虽然，皮亚杰理论对智力的看法比较偏狭，但这理论所假设

的智力成长阶段性则可有跨地区跨文化的效果。因此，在跨文化的智力研究中亦颇重视皮亚杰的智力测验。

三、资讯传处理论

资讯传处理论着重注释我们如何将一个问题的讯息传达到脑中，如何用脑中的短期记忆、运思、技巧以及如何在长期记忆中提取资料。进行哪些步骤。终而解决整个问题。简而言之，资讯传处理论是对智力活动所经历的心智过程的剖析去理解智力的性质。研究智力资讯传处过程有两个路向：一是着重个别智力功能传处速度。一是强调智力操作步骤的准确度和功能分析。资讯传处理论偏重心智运作过程的研究而不着重智力成果的考查。它关心的是在整个认知系统中，短期记忆的质素。思考推断的运用和长期记忆的资料如何影响人们整个智能操作过程。和如何影响人们在智力测验的表现。所以，这个理论的重点是了解智力活动的操作过程。一旦人们理解他们的大脑如何运送和处理智能资讯，才可着手加强思考的运用和智能的操作的训练。以下是一些类比试题研究的资料。

形象材料分类能力发展的研究[①]

对形象材料，例如图片等的分类研究有六种方式：

1、说出名称或理由。

先在儿童面前摆好正确分类的图片组，告诉他每组（类）名称，适当地说明理由，以此为范例，然后让其对实验材料

①朱智贤、林崇德著《思维发展心理学》，第198—200页，北京师大出版社，1988年4月第1版。

——各类事物的图片组说出名称或理由。

2、排除分类或归类

如下页图，在儿童面前摆好若干组（类）图片，每一组都有一张与该类无关的图片，要求儿童把这一张图片正确地挑出来。

3、直接分类或归类

例如，在儿童面前呈现没有固定次序的十二幅图画：鸡、鸭、杨树、白菜、大象、老虎、鹅、西红柿、茄子、狮子、松树、柳树，让儿童分类并说明种类名称及其理由，从中可看出儿童的归类水平。

4、指定分类或归类

将图片分成若干排（如第250页图中的三排），打乱图片的次序，让儿童从每一排抽出一个，然后分类或归类，并说明理由。

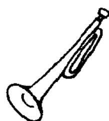
5、二级分类或归类。

将画有汽车、吉普车、自行车、火车、轮船、帆船、木船、步枪、冲锋枪、大炮、坦克、大刀、长矛、舢板、弓箭、宝剑等的十六张图片呈现在被试的前面，让他们归类，并说出名类名称。然后，让被试对分出的各类名称（词）再进行归类，这就是二级分类。如下表所示。

二级分类	一级分类	概念
交通工具	车辆	汽车、吉普车、自行车、火车。
	船只	轮船、舢板、帆船、木船。
武器	现代武器	步枪、冲锋枪、大炮、坦克。
	古代武器	大刀、长矛、弓箭、宝剑。

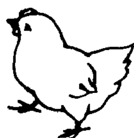
从分类的等级来分析被试的水平。

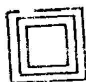
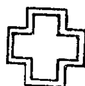
   

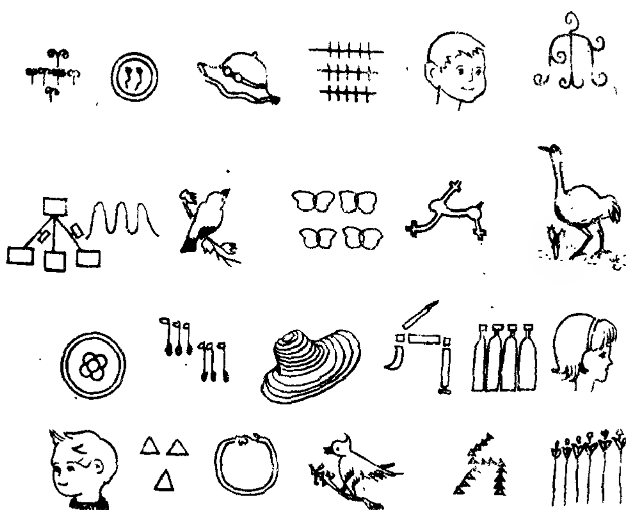
   



上述三种理论对智力持有不同的看法，但这只是着重点的不同而不是本质的不同。有的偏重智力运思的结构成份。有的强调运思形式的形式。亦有着重运思过程的步骤。虽然重点放在运思的结构，形式或过程三方面。三者皆认为抽象运思，逻辑推理，解决难题是主要的智能活动。同时，知识内容不同亦有其独特的规律也不同。须要一些特殊的逻辑运思去解决。因此我们不能创设一专门智能训练学科，均能解决各个不同学科领域的问题。智能技巧当然可以迁移，但不能全部由一学科转移到另一学科上应用。所以，智力测验成绩可预测一般学习成效，但其效度仍有局限，智力测验与学科测验都需要颇大部分的抽象思维和逻辑推理。但当这些运思过程用于不同学科时就削减了预测效度。其实，智力测验只能反映个人学习的总成果，不能评价个人的素质。智力测验成绩只能描述学生在这个特质中的个别差异情况。

第五节 智力测验结果的应用

智力测验并不是游戏。智力测验结果可以进行多方面应用，常见的有以下几种：

一、学习标准的确立

智力是学习的基本能力，只有了解学生的智力水平，才能确立学习进度，制定教学计划。学校的课程必须与儿童的能力相适应，这样就有必要对学生智力进行测定。所以，通过智力测验，能够对学生应注意的各个方面进行学习指导。

二、教育效果的诊断

智能指数一般是相对稳定的，但如果时而对学生进行特殊教育，时而较长时间任其自由放任，那么这种变化就会出现在智力测验的结果上。例如，比较间隔一定时间的二次测验结果，根据智能指数的显著增长和减少，能够判断教育的效果以及环境的影响。这种实验，在专门从事对学生进行连续观察工作的教育研究所和特别班级，尤为重要。智力测验在教育效果的诊断中起着重要作用。

三、学习评价的标准

根据五个阶段评价学生的学习，在学校具有一定的地位。但是，还有必要从儿童如何运用自己的能力，进行学习，开展学习评价，即在考虑智力和学习关系的基础上进行评价。一般说，智力和学习水平是一致的。但是在两者比例失调的

情况下，它有二种类型：一类是智力高于学习的几倍；另一类是智力低于学习的几倍。

四、应用于班级的编排

一般说，生理年龄的差异并不等于智力年龄发展的差距，众所周知，，幼儿和儿童（特别是低年级）在智力年龄发展上有很大差异。现代学校，一般以班级为单位，进行集体教学，而同一班级的儿童之间差异并不大，这样就容易形成表面化的高效率教学。因此按智力年龄为基础，编排不同层次的班级，比按照生理年龄编班更合理。

五、发现特殊儿童

区别低能儿童和超常儿童的智力问题。可以通过智力测验，但这不是单纯地判断智能指数，而是在测验过程中，通过观察被试者回答的方法，各种行为和态度，从而获得判断和指导这些儿童的有效线索。

与此相关，有些儿童到了学龄期，但智力年龄发展却极缓慢，以至不能接受小学正规教育，根据法律，可免去就学或延缓就学时间。在处理上述事情时，应在教育研究所和儿童研究所等机构，通过专家监定，获得正确的智力诊断

六、应用于升学，就职指导

让孩子从事何种职业，进入什么学校等事关前途的大问题，无论对教育还是家长都是很棘手的问题。要解决这些问题，应该采用智力测验进行诊断。现在社会上各种职业都有其不同的特点，一个人从事某种职业，并且干得得心应手，

这需要一定程度的智力水平。（如陆军普通分类测验AGCT.分数）。因此要将智力水平和职业性质结合起来进行升学和就业指导。

升学时，在听取个人志愿和家长要求之前，参考智力测验结果是大有益处的。要修完上一级学校课程，需要有一定的智力水平。如果达不到这个智力水平，勉强地升入上一级学校，其结果也是不理想的，对个人来说，就是不幸。如下就是这方面的统计资料：

IQ（智商）	
130	获博士学位人的平均智商数
120	大学毕业生的平均智商数
115	[illegible]
110	高中毕业平均智商数，[illegible]
105	能够修完高中智力课程可能性在50%的人
100	全国国民平均智商数
75	能读完高中课程可能性在50%的人

七、应用于性格观察

检查者和被检查者1对1进行个别测验时，通过仔细观察，能够获得判断被检查者性格的资料。智力测验成绩差的低能儿童，其智力的特性有很大差异，大致可分智力发展迟缓型和停止型。其中，智力发展停止型儿童，智力测验结果很差，伴有行动异常表现，测验成绩换算年龄不相符合，呈不规则性。智力测验中的试题一般按年龄大小、循序渐进，不断加

深，如果被检查者具有正常的性格和智力水平。在正常的环境下生活，那么，他的测验成绩不会呈不规则性，而大多数具有不规则性的儿童都或多或少有某种缺陷，这种缺陷有智能低下，感情脆弱，意志薄弱等。脑炎和脑膜炎患者的不规则性将更为明显。此外，经验告诉我们，幼儿的消化不良，也会阻碍儿童今后智力以展，但这种场合将不显示出不规则性，智力发展迟缓型多属于这一类。

第十章　考试的心理训练与卫生

考试的心理训练是指通过各种手段有意识地对应试者的心理过程和个性特征施加影响，使应试者学会调节自己的心理状态的各种方法，为更好地参加考试和争取优异的考试成绩，做好各种心理准备的训练过程。本章将对这一内容进行讨论。

第一节　心理训练概述

一、心理训练的概念

应试者在备考与实考中的心理训练有广义和狭义两种理解或解释：广义的是指在备考与实考中，对应试者进行有意识的影响，使其心理状态发生变化，达到最适宜的程度，以满足提高应试者的知识与智能水平和增强身心健康的需要；狭义的是指采用专门的仪器或手段，改变某一心理因素，以适应备考与实考的需要。两种心理训练各自有自己的目的、任务、方法与特征。也各自有自己的长短处。其比较如下：

首先，狭义的考试心理训练，是采用心理调节的专门技术手段进行训练；广义的考试心理训练则可以应用各种方法，不象狭义心理训练方法那样专门化。其次，狭义的心理训练，要求提高具体的心理素质或克服某种心理障碍。如改善注意强度或克服情绪的紧张等等；而广义的心理训练则着眼于心

理状态的普遍适应的改善，一般并不要求针对某一种心理因素进行训练。再次，狭义的心理训练要求具有明显而快速的心理训练效果。例如，用心理训练的方法恢复应试者脑力与考试后的自然恢复是不同的，而我们见到通过恢复心理训练的应试者是精力充沛，感觉敏锐等。而广义的心理训练由于涉及到的问题较多，一般短期内不易看到直接的效果，有些效果也往往是无形的。

当然，考试中的两种心理训练是紧密结合的，互相补充的。只有广义的心理训练，缺乏针对性和具体手段，不容易看到实际效果，不利于坚持长久；只进行狭义的心理训练，缺乏全面的心理训练的基础，不利从根本上改善应试者心理状态。作为统一的心理训练概念，不应当人为把两者割裂开来。否则就会违反辩证法的基本思想。

二、心理训练的地位与作用

（一）心理训练的地位，心理训练是参加现代考试的重要组成部分，它和知识学习（或温习），智力的开发与训练等一起构成了备考的基本内容。科学研究表明：现代考试要求应试者不仅要消耗生理能量，而且还要付出巨大的心理能量，因为现代考试的备考与实考在应试者的机体施加生理负荷的同时，也施加了心理负荷。应试者没有良好的心理准备状态，就不能顺利地完成备考与实考。更难以夺取优异的考试成绩。那种认为只要平时学习与温习得好，实考就会自然出好成绩，或者平时不用下功夫，考试时凭“一股劲”就能拿好成绩的看法，都是不合符现代考试发展规律的。从国内外一些考试成功者可以看到，在认真学习与复习的基础上，

心理因素往往在考试的成败上起着重大的作用。

（二）心理训练的作用在于发展应试者参加考试所需的心理品质，使其对备考与实考具有心理准备性和稳定性，其作用主要在于使应试者或学生的各种心理过程和个性心理特征更快地得到完善和发展，形成备考与实考的最佳心理状态，从而帮助应试者顺利完成备考的任务和取得优异的考试成绩。

首先，促进应试者心理过程的完善。

人的心理过程包括认识过程、情感过程和意志过程三个方面。极度紧张的备考与实考对应试者的心理状态提出了更高的要求。知识水平，智力高低和心理素质（专项考试心理因素）是决定应试者的备考与实考成败的三个不可分割的因素，其中知识水平、智能高低，是保证应试者考试质量的物质基础，基本条件。而心理素质是使两者能够发挥作用的内部动力，对应试者来讲，心理因素是他们控制自己的生理活动和智能发挥的主导因素。实际心理活动水平太低，不能进行有效的控制，在这种情况下，尽管具有较好的知识水平与较高的智力，也不能使其充分发挥作用。甚至有时越是知识水平好，智能水平高的人，反而失常得更厉害，这是因为若没有心理训练，使知识、智能水平高的能量，冲击应试者的心理状态，产生心理紧张，使回忆、思维等出差错，或二者脱离有机的联系，等等。为此，必须用心理训练的方法，提高应试者心理活动的强度，以达到能进行自我控制的水平。

其次，促进应试者的个性心理特征的形成与适应考试心理状态的形成。

人的个性心理特征包括性格、气质、能力、兴趣、动机

等方面。在备考的极度紧张的条件下，决定应试者或学生行为特点的最重要的个性心理特征是应试者的动机。对备考和实考的兴趣程度、个人的性格特征和气质等。心理训练可对应试者的良好性格的形成发展产生巨大影响，可以改善人的兴趣品质，可发展或改变气质的某些特征。可促进考试必须的特殊能力。

应试者的心理状态是最容易变化的心理结构，它是考试所必须的最重要的心理机能的综合表现，其特点是有一定的积极性和强度。心理状态的特点和水平对考试的进行与结果有很大的影响，考试中取得的成绩和自身的各方面的提高，在很大程度上取决于对应试者心理状态的控制和自我控制。心理训练有助于培养应试者心理过程的稳定性，发展在极端紧张的活动时控制自己心理状态的能力，形成参加备考和实考的适宜心理状态。

再次，消除心理障碍的进程得以加快。

考试心理训练的作用，不只是限于对心理活动水平的提高或降低的调节，还有消除和医治某些已经形成的心理障碍的作用。

在备考与实考中，由于在某一考试上的失误，往往会造成心理上的障碍。如在高考中临场情绪过敏，心理疲劳、动机不足，大脑反应迟钝等等。这些心理障碍是由于考试挫折直接引起的心理伤痕。对此，一般需要采取专门性的心理恢复或治疗措施，不能用只对考试内容（知识、智能）学习的方法来代替。心理障碍只能用心理学的方法去克服。科学研究的材料表明：克服心理障碍，只能采用心理训练的方法，不能依靠知识的学习，智能的提高来代替，也不能放弃修复

性治疗，单纯依靠自然恢复。有些较轻心理障碍，如心理疲劳，可以通过自然恢复消除，但有些心理障碍却不能完全依靠自然的恢复，必须进行专门性的心理训练。自然恢复不仅不能根治某些心理障碍，反而可能使其加重，造成习惯性的心理过敏。

总之，考试心理训练是为了培养应试者具有适应备考与实考所需要的多种心理品质和心理能力；加速应试者对应考内容的掌握与提高的熟练程度。形成应试者对备考与考试的良好态度，创造适宜的心理状态，提高适应考试的能力；克服应试者的各种心理障碍，以保证备考与实考的顺利进行，促进应试者的疲劳的恢复。

三、心理训练的分类与任务

考试心理训练的分类根据不同的训练内容和参加人员的情况，可分为一般心理训练和准备具体考试的心理训练。一般心理训练在备考与实考期间都可进行，准备具体考试训练则一般在考试前一段时间开始进行，并一直持续到实考期间。

（一）一般心理训练与任务

一般心理训练是在长期的训练过程中培养和发展应试者应试所必须的各种心理品质和心理能力的训练过程。一般心理训练的具体任务包括：

1. 培养应试者对考试的兴趣、能力、性格、气质等个性心理特征。

2. 发展考试所需的知觉、记忆、表象、想象、形象思维、抽象思维及情感和意志品质等心理过程。

3. 发展注意品质，包括注意的稳定性、注意集中、注

意范围、注意转移、注意分配等。

（二）准备具体考试的心理训练与任务

准备具体考试的心理训练即短期心理训练，它是在较短的时期内使应试者学会自我调节心理状态的方法，加速形成考前最佳心理状态的训练过程，其任务包括：

1. 使应试者明确考试任务，激发良好的考试动机，建立取胜心理定向，形成实现目的的信心。

2. 使应试者掌握各种具体心理训练方法来调节和控制自己的心理状态，消除紧张情绪和心理障碍，形成最佳考试心理状态。

3. 使应试者学会在千变万化的考试情况下保持积极稳定的心理状态，顺利完成复杂而艰巨的考试任务。

准备具体考试的心理训练又包括考前和实考过程中的心理训练

考前心理训练一般在考前两三周开始，可根据考试具体任务。对考试的水平，场地气候条件，以及主考者（考试机构）的意图进行安排，通过考前心理训练，可使应试者具有良好的考试动机，消除紧张情绪，增强取胜信心。

实考过程中的心理训练可在每次考试前，一次实考中，两次考试间以及考试后进行。它是帮助应试者分析在实考过程中出现的新情况，积累经验，及时修订备考计划，根据考试中应试者的心理表现采取必要措施。使应试者在整个实考过程中保持稳定心理状态，顺利完成考试任务。

四、心理训练的原则

心理训练不同于知识学习与智能提高的训练，它有其自

身的特殊内容与方法。因而，在心理训练中除了必须遵循知识学习与智能提高的训练的一般原则外，还必须注重心理训练自身的独特性。

（一）必须促进应试者的心理发展

心理训练是对应试者的心理施加影响的训练，它是直接转化人的“内心世界”的特殊教育过程，心理现象人们称之为“心灵”，是人身上最宝贵的部分，也是最容易变化和损伤的部分。心理训练作为考试学的一个新的内容和分支，是有自己独特性的，它的训练对象是人的精神活动，即人的“心灵”，“心灵”是统帅整个机体，影响全部行动的重要部分。为此，任何心理训练方法的使用，必须首先有利于应试者的身心健康，促进考试成绩的提高。

根据这一原则，在进行心理训练时，不能采用消极的训练方法，即不能因心理训练而给应试者带来身心痛苦和损伤，对待心理训练，不能象做其他自然科学实验那样，采取先“损坏”，再恢复的办法来进行心理训练。例如，不能用冷落、讥讽等粗鲁言行来“激将”，激起考试的“斗志”等。假若在个别人身上能收到一些暂时的“效果”，然而它留下的心理伤痕是很难修复的，从根本原则上讲，有意识地对人施以不利的心理影响，从而获得某些表面效果，是不合符道德行为的。心理训练必须坚持科学的人道原则，维护应试者的身心健康和尊严，要从关心，爱护发展应试者的身心的立场出发，这是由心理训练这门新兴学科的特殊研究对象所决定的。

心理训练是一项细致的科学试验工作，是塑造“心灵”的实验，事先必须对被试对象进行有关科学知识的教育，要

有周密的计划步骤和科学的心理训练修养。决不能粗心大意，掉以轻心。

（二）必须把应试者个体特点与考试实际相结合起来。

采取个体化的心理训练方法，主要是考虑每个人的心理潜力，弥补心理欠缺，根据个体特点促使心理的均衡协调发展。在心理训练时不仅要以个体心理特点为依据，而且要以每个个体在不同时间内的具体心理状况的变化为依据，人的心理是有较大的可塑性，这是人所共知的，进行心理训练不能不考虑这些特点。

根据个体心理特征进行心理训练时必须考虑应试者所考科目。不同的科目，要求不同的个性特征，而事实上，进入各个科目考试的人，其个性不一定完全符合该科目的要求。在进行心理训练时，必须把个体[illegible]科目的[illegible]求，使每一个参加心理训练的人，充分发展某些特殊的心理素质，克服某些心理欠缺，以满足他们备考的需要。

考试心理训练的直接目的，在于促进知识的提高和智能的发展。使其获得优异的考试成绩，一切考试心理训练的运用，都应在备考和实考中进行。为此，尽量使考试心理训练在结合考试项目的实际前提条件下进行是完全必要的，另一方面，考试心理训练在于改善心理状态，使其达到最佳水平。以适应考试的要求。而改善心理状态必须以应试者的个体身心特征为依据。因而，心理训练绝对不能千篇一律。必须针对不同的个体的身心特点和实考项目的具体情况进行心理调解训练。这样，不仅可以加快心理训练的应用过程，而且还可以丰富和改进心理训练的方法，使其注意心理训练的调节和控制作用，不致因忽视心理因素造成备考与实考失利。

（三）必须坚持自愿与持之以恒相结合。

心理训练是应试者自我调节心理状态的训练，主要手段都是由应试者自己掌握的。因此，被训练者能否志愿配合，是心理训练效果好坏的主要因素，当然，强调自愿，决非没有较好的外界诱导因素。而是建立在应试者的需要的基础上的，应试者进行心理训练时，如果采取积极态度，就会很快地掌握自我调解之手段，取得预期的效果；如果他们对此持观望、怀疑，甚至否定的态度，这就会成为自我调节方法的内部心理阻力，增加应试者的心理负担。心理训练要求从根本上改变人的心理状态和个性特征，这不是轻而易举的事情，必须动员受训练者具有耐心和信心，持之以恒，不断进行自觉的自我训练。而且要准备在心理训练过程中，经过波折，逐步学会控制自己的心理状态，急于求成不仅无益于自我调节，反而欲速则不达。所以考试的心理训练必须坚持持之以恒的原则。

第二节　心理训练的基础

心理训练理论和方法来源于两个方面，一是来源于印度的瑜伽和中国的气功，它是以养身为目的的自我锻炼方法，后来开始在欧洲用于对病人的心理治疗方面；二是来源于心理学的实验方法。两者结合起来并运用于考试的心理训练方面是它的物质基础。

一、心理训练的气功基础

我国最早的一部医学巨著《黄帝内经》中指出："药不

能独治”，“恬淡虚无，真气从之，精神内守，病安从来？”这里说的是治病不能单用药物，而且要用气功方法进行引导，使人减轻精神负担，做到“恬淡虚无”，“精神内守”，不胡思乱想。能控制自己集中注意力，这样，就可以达到精神治疗的目的。“导引行气”的养身治病的“气功”训练方法来源于此。气功把修心养性，解除人的精神负担，作为第一个基本方法，即意守功。意守功有一系列对意识进行自我控制的手段，运用这些手段，使人的意识获得主动自守，而意识的安定又会给整个身心的安静带来好处，使人养精蓄锐，恢复身心能量，为健身的工作打下良好的心理及物质基础。

在考试的心理训练中，训练注意集中是首要的问题。因为不少人正是由于临场的注意力分散而引起情绪紧张，才失去对心理的控制。心理训练中的注意集中训练和气功的意守功的基本思想相同，而且前者是后者的借用和发展。因此，气功的意守功是注意集中训练方法的基础。

气功的放松功能使应试者在紧张的备考或实考后进行放松。气功对肌肉，骨骼关节的放松练习叫做“放松功”，它要求练习者，从肌肉到骨关节，从外部感官到大脑皮层都要逐渐放松，通过具体的放松动作，使整个机体和心理活动都处于放松状态。这种放松功正是考试心理训练的来源和基础。

我国出土文物《行气玉佩铭》中记载着我国公元前约380年的战国初期，气功吐纳调息的练功要领，其中刻画的“吹嘘（Xu），呼吸”、“吐故纳新”方法为：“行气深则蓄，蓄则伸，伸则下，下则定，定则固，固则萌，萌则长，长则退，退则天，天几舂（Chōng）在上，地几舂在下，顺则生，逆则死。”这段话讲的是调节呼吸的规律。大意为：吸气能

深入则能多其量，量多了往下伸延，气伸延到下部就能感到安定充实，气稳固后再呼出时，则如草木之萌芽，往上长，这时与深入吸气的路线相反要呼尽，一直呼到顶，这样便会带动整个机体上下运动，按照这种呼吸方法就能增强生命力，违反它就会导致疾病或接近死亡。这段话概括了气功调息功的基本要领，限于当时历史条件，有些论述过于极端，也是可以理解的，但是"调息功"（又称吐纳功）在心理训练中的调节呼吸方面很有用处。在调节呼吸方面基本上是借用"调息功"的经验与方法。

气功的方法很多，在心理训练上的运用也是千姿百态的。上述几种已足够说明心理训练与气功之间的紧密关系了。印度佛教的"瑜珈"术与我国气功练习很相似，也是一种调节身心的训练方法。"瑜珈"一词出自印度的一种教义，是梵文的音译，意思是"上挽具"，"将马拴在辕上"，它和"训练"一词相近。瑜珈练习的基本内容也是肌肉骨骼的放松，呼吸调节和类似意守的"冥想"等等，主要是通过瑜珈练习，使人达到身心调合统一。控制身体的情欲，解脱人的精神痛苦。

二、心理训练的实验心理学的基础

德国心理学家、哲学家、构造心理学派创始人冯特，1879年在莱比锡建立世界上第一个心理实验室、采用自然科学的实验方法，对人的心理进行实验。从而，把心理学从哲学中分化出来，使其成为一门独立的新科学。他使心理现象的研究走上了客观实验的道路。实验心理学借助于自然科学，主要是用生理科学的手段，把复杂的心理现象在仪器的控制下

呈现出来，给以规定的刺激，引起特定的心理现象并精确地记录下来。考试心理训练，应用类似心理实验的手段，用特定的仪器，给予某种信息，使应试者按照指示信息进行某种反应（动作或语言等），然后将这些标志心理变化的反应记录下来，经过反复练习使应试者学会主动引起并控制心理现象。这就是现代心理训练的方法。例如，进行Toefl考试时，可用仪器设备将自己在考试过程中的思维、记忆、解答问题时的显象记述下来，认真研究，以检查考试中各类题目中出现的各类问题。提供给下次考试时参考。在借助仪器的帮助下进行训练，应试者通过仪器的帮助，逐步学会对某种心理现象的自我调节能力，并达到脱离仪器独立自我调节心理活动的程度。这就是考试心理训练的科学方法。

当然，作为心理训练基础的心理实验，不等于就是心理训练。心理实验主要是运用仪器引起某种心理现象、精确记录和加以客观说明，并不要求对这一心理现象进行重复训练，而且为了获得实验的客观数据，还要控制应试者的练习因素；而心理训练则不同，它不仅仅借助于仪器引起某种心理现象并记录它的效果，而且还要求不断重复练习，使其达到熟练程度，应试者要学会自己控制这种心理现象。而心理实验则要求尽量避免应试者的主动参与，可见两者是有原则性区别的。一个只要求验证心理现象，另一则是要改变并且学会控制这种心理现象。两者的共同性在于都是使用实验心理学的科学手段，使实验或训练科学化、客观化。除此之外，心理训练在某些方面还可以采用其他方法，如谈话法，自我暗示法等等，不一定都必须借助实验仪器进行。所以实验心理学是心理训练的基础之一，但不是唯一的基础。

总之，心理训练是考试心理发展中的新的阶段，它的出现不是凭空的，是现代科学技术发展的成就，它以丰富的考试心理，尤其是考试的经验为基础，它将作为一门多基础的应用心理学科的分支成长起来。

第三节　应试者的心理概述

众所周知，各类考试（特别是升学考试）牵动着千家万户的心，在内外压力下的应试者们在考前有哪些心理表现呢？人们在这方面要注意哪些问题？这是一个较为深刻的问题。曾经有人对469名中小学毕业生考前心理调查[①]，采用问卷法得出考试前部分心理表现，如下页表1，表2所

①孙逊：《469名中小学毕业生考前心理调查》，载《齐齐哈尔师范学院学报》1985年3期。

表1：中小学毕业升学考试前部分心理表现

人数 \ 年级			小学	初中	高中	合计
		人数	74	273	122	469
对待升学	是自己唯一出路	人数	72	173	50	295
		%	97.30	63.37	40.98	62.90
对待不能升学	无所谓	人数	—	53	48	101
		%	—	19.41	39.34	21.54
	怕老师，同学邻居瞧不起	人数	42	84	20	146
		%	56.76	53.16	16.39	31.13
	怕父母责骂	人数	10	29	2	41
		%	13.51	10.62	1.64	8.74
对升学考试	充满信心	人数	29	121	38	188
		%	39.19	44.32	31.15	40.09
	信心不足	人数	39	135	82	256
		%	52.70	49.45	67.21	54.58
精神压力	大	人数	43	121	43	207
		%	58.11	44.32	35.25	44.14
	中	人数	19	104	48	171
		%	25.68	38.10	39.34	36.46
	小	人数	6	43	26	75
		%	8.11	15.75	21.31	15.99
平时考试是否	紧张	人数	29	159	75	263
		%	39.19	58.24	61.48	56.08
	不紧张	人数	2	44	27	73
		%	2.70	16.12	22.13	15.57
考前复习	慌乱	人数	56	180	92	328
		%	75.68	65.93	75.41	69.94
	灵活	人数	17	88	31	136
		%	22.97	32.23	25.41	29.0

表2：父母、教师的态度对考生心理的影响

学习情况	调查人数	性格孤僻或暴躁（%）	平时学习成绩差（%）	精神状态昏头昏脑%
训斥打骂	10	7（70.9）	5（50.0）	6（60.0）
关心帮助	59	12（20.34）	8（13.36）	15（37.28）
X^2		10.57	7.43	4.82
P		<0.01	<0.01	<0.05

表3：平时学习成绩与毕业生考前心理

学习情况		学习较好	学习较差	X^2	P
调查人数		48	88		
升学为唯一出路	人数（%）	34（70.83）	41（46.59）	7.38	<0.01
有个工作就行	人数（%）	4（11.76）	27（30.68）	8.81	<0.01
自学可以成材	人数（%）	21（43.75）	18（20.45）	8.24	<0.01
精神压力大	人数（%）	10（20.83）	57（64.77）	23.99	<0.01
迎考信心不足	人数（%）	12（25.0）	85（96.59）	77.82	<0.01
怕讽刺瞧不起	人数（%）	4（8.33）	31（35.23）	11.75	<0.01
考前复习慌乱	人数（%）	21（43.75）	84（95.45）	47.18	<0.01

表4：中小学毕业生考前精神不良状态情况（%）

年级 \ 性别	小学		初中		高中	
	男	女	男	女	男	女
昏头昏脑	29.03	44.19	23.21	33.54	12.70	22.03
合计	37.21		29.30		17.21	

上述4个表的数据充分说明，应试者考前的心理状态与平时成绩和应试者的身心健康、父母、教师的影响等有极其紧密的联系。为此，本节将对应试者的最佳心理结构进行探讨。

一、应试者的最佳考试状态

经常有考生说自己今天考试没有达到“最佳状态”。但是，到底什么是考试状态，人们的回答是不相同的。对考试状态，有人理解为备考状态，有人理解为身体状态，有人理解为心理状态，也有人说是备考状态与实考状态等等。这些不同的理解都对这一概念的涵义表述不准确。因而在实际考试中造成误解。“考试状态”是指实考和备考两种状态，其中包括知识水平、智能高低、身体素质和心理素质三个方面的内容，它是一个多因素的综合概念。“实考状态”是指实际考试实施时的知识、智能、身体和心理状态。有些人正是由于考试的状态与实考状态混淆起来，误认为只要有了平时复习时的知识、智能水平，实考时就能出好成绩。其实并不尽然。不少人平时复习时的知识、智能水平很高，但是到考场上就发挥不出来了。可见最佳考试状态不同于最佳实考状态。考试状态包括平时复习状态，应试者的备考应当包括考前的知识复习和应试的心理训练两个方面。在平时复习中，没有最佳心理状态的保证，应试者打不好基础，关键性的知识掌握不住，光靠实考时的临场努力，取得好成绩是不大可能的。只有最佳备考状态，没有考试时的最佳实考状态，再好的平时备考，也不能发挥出来。区别最佳考试状态和最佳实考状态，并且采取不同的备考与训练准备进行有计划的平

时和实考前准备，这是十分重要的课题。那种认为只要平时复习好，实考时就会自然出成绩，或者平时不用下功夫，实考时凭“一股劲”就能考好成绩的看法，都是不妥当的。

二、应试者的最佳心理状态

应试者的最佳心理状态是最佳考试状态的一部分。它分为实考和备考最佳心理状态两个方面。由于应试者的平时备考的考前的心理训练与实考的条件不同，要求的心理因素和心理水平也就不一样。为了使应试者在考试中取得好成绩，必须研究这两种最佳心理状态组成的具体心理因素以及它们的发展水平。为了把问题说得比较集中些，我们不妨把两者放在一起来讲，而以对实考中的最佳心理状态的分析为主。

不少人对于最佳心理状态有一种简单的理解，即认为最佳心理状态，似乎只是一种单一的心理因素。如象，当考试成绩很不好时，总是说我的情绪不好或我的注意力不集中等等，当考试取得较好的成绩时，则说实考时觉得很敏锐，思维很清楚。有的说注意、情绪，有的说感觉、思维。到底一个应考者的实考与备考的最佳心理状态应当包括哪些心理因素，许多人并不清楚。对最佳心理状态的因素不清楚，就不能准确而全面地分析一个应试者的心理状态。至于在实考中，特别是在关键性的时刻，使应试者取得了优异成绩的有哪些心理因素达到最佳状态。甚至转败为胜的关键性的心理因素是什么，他们就说不准了。但是，造成考试失败心理因素是什么？人们的看法往往也不一致了，有的认为是情绪问题，有的认为是意志品质问题，也有的认为是注意力的问题，各说不一。最佳心理状态的许多心理因素，绝大部分是不能看

到的，能够看到的只是它们的外部表现；如注意、情绪等。心理因素是一种内部状态，在一定时间内并不外露，单纯凭外部表现做直观的判断，是不科学的。例如，考试成功时，有利的心理因素显露出来，而不利的心理因素掩盖起来；失败时，有利的心理因素掩盖起来，不利的心理因素显露出来。这就造成两种情况：考试成功，掩盖的不利心理因素成为潜在危险，一旦遇到适当时机就会显现出来；失败后，有利的心理因素被掩盖，又有失去其作用的危险。可见，正确分析应试者最佳心理状态的组成结构，找出它们形成条件和个体差异，是使应试者具有最佳心理状态的重要问题。

应试者的最佳心理状态是由许多心理因素组成的复杂心理结构，它至少应当包括下列几方面的心理因素：信心、意志、情感、注意、思维等。当然，这不是全部因素，还有记忆、个性等心理因素。

三、应试者最佳心理状态形成的基本因素

（一）充足的信心和顽强的意志

充足的信心和顽强的意志是构成应试者最佳心理状态第一个重要的心理因素，充足的信心和顽强的意志，可以充分调动一切有利的心理素质，使应试者处于积极的、顽强的战斗状态；信心和意志是内部的精神力量，可以调动一切有利于考试的知识、技能与技巧，使其得到充分的发挥。因而信心和意志又是应试者最佳心理状态中核心的心理因素，是起决定作用的主导因素。

应试者的信心是考前对自己潜力和考试的正确认识。这种认识越准确地反映其实际情况，那就越易取得考试的成功

应试者的自身潜在条件好，备考充分，超过了客观不利条件，它就会产生充足的信心。反之，就会使信心不足，人们经常关心的是信心的大小问题，但是很少注意信心形成的客观基础问题。如果缺少对客观基础的科学分析，信心的正确性就会降低，信心的性质就会改变，会由积极的性质变为消极的性质。一个应试者脱离了自身和客观条件所建立的“必胜”信心，往往会夸大自己的潜力，轻视客观的因素，在需要应付困难局面时反而缺乏信心，束手无策；反之，离开内外基础的信心不足，将会在有利的客观形势下，不敢去争取考试的成功。而限制自己的内在潜力的发挥，在考试之后感到精力有余，为未能充分调动自己的力量而遗撼。

根据有无充足信心可以将考试信心概括为四种不同的类型。

1. 充足信心的积极型。这种类型的应试者的信心力量很大。他的信心是建立在自身和外界条件基础上的，所以，它能最大限度地调动应试者的自身的内在潜力，克服自身和客观条件的不利因素，使心理素质，知识智能的优势充分发挥作用。这种信心类型，是出考试优异成绩的类型。

2. 充足信心的消极型。这种类型的应试者对考试有充足的信心。但，这种信心不是建立在对自身及外界条件正确认识基础上的。往往以过高估计自己力量，或过低估计考试为基础，以致在紧急的场合，使信心发生消极作用，在这种情况下，信心强度越大，起的反作用越大，甚至导致考试失败。

3. 信心不足的积极型。这种类型应试者的信心基本上是建筑在对自身条件和客观形势正确认识基础上的，只是由

于某些因素或其它原因，使信心的力量较弱。这种类型应试者的能力也能发挥到一定程度，但不能发挥到最高水平，更不能超水平地发挥。这种信心类型是那些保守，稳重的应试者经常表现的心理状态。虽然他们的信心强度不大，但是能起积极作用，当然积极作用是有限的。

4. 信心不足的消极型。这种类型的应试者不仅信心的强度不足，而且态度消极，过低估计自己力量，或过高估计客观困难，同时又缺乏改变不利现状的积极态度，因而限制了自己的考试时所应具有的能力与水平，长期处于无所作为的状态。对这种应试者来讲，最主要是信心消极性的问题。

（二）稳定而灵活的注意

注意是应试者进行心理训练的重要因素，它是构成应试者最佳心理状态的组成部分。

所谓注意即心理活动的指向性和集中性。指向性是对一定事物的选择；集中性是对所选择的事物的贯注和坚持。这种选择、贯注和坚持的积极状态，就使人脑能够清晰地反映周围现实中的一定事物；而对其它事物则反映得模糊不清，甚至完全没有反映。注意本身并不是一种独立的心理过程，它是各种心理过程的一种共同特性。每一种心理过程都总是程度不同地指向和集中于一定的对象的。注意在人的实践中起着很重要的作用。引起注意的原因有时是事物本身的特点，如强烈，新奇，对比明显、不断变化等客观因素，有时是人的主观因素，如当前的任务和态度，一般精神状态，以及个人的兴趣、需要、知识经验和世界观等。在一定条件下，主观因素对选择对象和维持注意有决定性作用。注意是一种“定向反射”和随之发生的“适应性反射”。它的生理机制

主要是大脑皮层的优势兴奋中心和相互诱导的作用。“定向反射”发生时，大脑皮层的一定区域就产生优势兴奋中心，并由此产生机体的各种适应性反射活动，因而大脑就能够更清晰地反映有关的事物。与此同时，皮层的其余区域由于负诱导的作用就处于一定程度的抑制状态，而与这些区域相应的事物就不能引起反应，或不能引起清晰的反应。这种“定向反射”和“适应性反射”与人的第二信号系统的活动具有紧密的联系。

但是对于应试者来讲，他所注意的对象是多是少，注意集中的时间是长是短，要根据不同的考试内容，根据考场的情况来决定的。因此，需要针对不同具体考试内部以及应试者个性特征来训练最佳注意状态。应试者要取得好成绩必须保持注意的稳定性，又必须保持注意的灵活性。

注意的强度是指应试者的意识对一定对象指向和集中的程度，指向的范围越小，集中的时间越长，注意力强度就越大。反之，如果注意很快就离开对象，或集中的时间不长就疲劳，都是注意强度不足的表现。注意集中的强度过大，不仅影响思维与回忆的准确性和协调性，而且也不利于其他心理因素发挥作用。注意太强可以使情绪紧张，思维片面化，也可能产生错误的知觉等。

为了使注意成为考试最佳心理因素，只是强调它的强度是不够的，同时还必须强调注意的灵活性。注意没有灵活性，就无法分配或转移，这对备考和实考都是不利的，任何考试都包含着许多方面，而各个方面又都依赖于一些条件的变化。为了使多个方面能取得好的成绩，注意必须合理分配和转移，做到对各个方面的统筹兼顾。

（三）充沛而稳定的情感

考试的情感是应试者对备考实考活动的态度体验。这种体验是对事物与个体需要之间关系的反映。当外界事物满足应试者的个体需要时，他就产生愉快、积极的考试情感；当外界事物不满足，甚至破坏了应试者个体考试需要时，他就会产生不愉快、悲伤或愤怒的考试情感，应试者的需要不仅限于考试本身的需要，还有其它的社会和生理的需要。但一个应试者情感只有建筑在正确的考试需要的基础上，才可能是充沛而又稳定的。充沛而稳定的情感是应试者考取优良成绩，克服各种困难的力量源泉之一。应试者对考试不能离开自己的情感的鼓动，无动于衷，消极淡漠是不会取得很好的成绩的。

造成应试者情感强度大小的原因，是多方面的，除了气质类型特点以外，还有社会的，家庭的以及个人动机，兴趣等心理方面的原因。但是经常起作用的，最直接的原因乃是他们对考试的需要，一个为了进入大学学习，而把进一步深造作为自己的一切希望的人，他的情感一定是充沛的。

情感作为最佳心理状态的组成因素，除应具有积极性和强度的特点之外，还必须具有稳定性的特点。情感的稳定性是指情感产生之后维持的时间长短，情感的稳定性是和情感的强度分不开的，一般讲，越是强烈的情感，越比较能稳定。但是，有些人具有强烈的情感而不稳定，在激烈的考试争夺中最需要充沛而稳定的情感。不少应试者在考试顺利时候，容易表现出情感的强度和稳定性，而且容易表现出积极性的情感状态；在失利时，则情感的强度会降低，也不能保持情感的稳定。

第四节　考试心理训练的基本方法

考试心理训练是多方面的，它包括认识、情感、意志等心理过程的训练和个性特征的训练，以及记忆和注意的训练等等，凡是对某种心理现象施加影响，使其发生变化的措施，都可称为心理训练，在考试心理训练中，由于考试科目的不同，从事各科目的应试者的个体心理特点也不同，选择对他们进行心理训练的内容，也应当有所不同，根据第一节分类，有一般训练方法和准备具体考试的心理训练方法。

一、一般心理训练方法

一般考试心理训练主要是改善心理品质。改善考试训练所需的心理品质的训练方法用得较多的是注意集中训练和意志品质训练。

（一）注意集中训练

注意集中是坚持全神贯注于一个确定目标，不为其它念头所分散的一种能力。注意集中训练就是阻断自己的思想与其它无关事物的联系，把注意指向和集中在特定的事物之上的练习方法。

注意集中训练可采用以下方法：

1.听音法：把歌曲或英语磁带用录音机放出微弱的声音练习听觉（时间3－5分钟为一段），这种可以间接地提高应试者的自觉集中注意力。

2.视物法：选择一个目的物（任何一件物体），对其仔细观察几秒钟后，闭上眼睛努力回忆被观察物的形象。如

果回忆的物体形象不太清楚，就睁开眼睛再看一遍然后再闭上眼睛回忆。如此重复数遍，直到在头脑中清晰地回忆出被观察物体的形象为止。

3. 看表法：注视手表秒针的转动。先测试最初所能持续的时间，记录后进行练习。每次练习重复3--4次，间隔10—15秒。注视的时间可以从60秒至90秒至180秒逐渐增加，能持续注视5分钟不转移注意力就算不错了。最好白天练一次晚上临睡前练一次。

（二）意志训练

意志是人们为了达到既定目的而在行动上所表现出来的自觉克服困难的心理过程。意志品质是意志的具体表现，包括自觉性、主动性、勇敢性、顽强性、果断性、自制力等。

意志训练是有目的地使应试者克服困难，完成意志行动，从而提高意志品质的过程。

培养应试者意志品质的途径主要有：

1. 努力克服困难。人的意志总是与克服困难相联系的，即在克服困难的过程中表现出来的。考试训练中遇到的困难可分为客观困难和主观困难两类。

客观困难是同考试项目特有的障碍相联系的困难，如数学难题。英语单词的遗忘等。主观困难（又称为心理困难或心理障碍）是应试者对待备考和实考的条件所抱的主观态度方面的困难。如因高考落榜而产生的恐惧，在参加新的考试科目时的害怕等等。

克服客观困难所必需的意志品质是直接在考试复习过程中培养的。而克服主观困难就是心理的问题，在克服主观困难的过程中，应试者的勇敢与顽强精神可得到发展。

2. 掌握心理自我调节方法，应试者要善于控制自己的思想情感和行为，即具有自制能力，才能在紧张的考试中充分发挥已经获得的考试能力。自制性以情感的稳定性为基础，因此在训练中要掌握心理自我调节的方法，善于控制自己的情感，保持适宜的情绪兴奋程度。即使在内部状态不良的情况下也能充分地发挥原有的水平。心理调节的方法有自我鼓励、自我说服、自我命令及自我暗示和放松训练等。

3. 学会自我培养意志能力。意志品质的培养在很大程度上取决于应试者的积极主动性。因此，应试者要养成自我培养意志的习惯，首先应试者要明确考试的目的任务和意义，目的明确可激发应试者克服困难的自觉性和主动性，在备考和实考中表现出坚强的意志以达到既定的目的。其次应试者要了解意志培养的原则与方法，学会对自己的意志品质进行分析，结合考试项目实际有针对性地进行培养磨炼，并努力克服自己意志品质薄弱的方面。

二、准备具体考试的心理训练方法

（一）考前心理训练方法

1. 模拟考试训练。模拟考试训练是用接近考试的实际情况对应试者进行实战的反复练习，以提高应试者对考试的适应能力的心理训练方法。模拟训练可预防应试者实考时不良心理状态的发生，提高他们的心理稳定性和应变能力。

2. 情绪控制训练。情绪控制训练是采用一定手段有意识地调整应试者考前不良情绪状态的心理训练方法。考前不良情绪状态有考前过分激动状态，考前淡漠状态，考前盲目自信状态等。这些消极情绪会导致应试者的注意力涣散，考

试能力明显下降，不能在实考中发挥原有的水平。

调整考前不良情绪的方法有：有意地改变表情动作，不同节奏的呼吸练习，逐级使大脑放松，积极正面的语言暗示，改变影响应试者情绪的条件等。另外还可以通过生物反馈训练，模拟训练和练习考试来使应试者学会控制情绪状态的方法。

进行情绪状态控制训练要考虑应试者的个性特征，如应试者的个性、性格、气质、意志品质、情感特点等。

（二）实考过程中的心理训练方法

1. 直接心理训练。

是指直接在完成考试行为时进行的心理训练，它是在考试和注意紧张的条件下进行的。直接心理训练包括检查定向、集中、评价三个环节。

检查定向是从应试者来到考场起到实考开始，要求应试者进行全面注意观察和直观有效的思维，并采取具体的实际行动，如检查一下钢笔的墨水，回忆一下几个简单的公式或定理等。集中是排除所有无关思想和外部刺激，把注意力完全集中到将要完成的考试中。已做好实考的充分准备是这一环节结束的标志。应抓住这一时机马上开始实考。在完成一次实考后，应对完成情况进行评价。评价主要靠回忆来进行，需要时可在下次实考中做必要的行为修正。

2. 自我暗示和放松训练

自我暗示和放松训练是以一定的套语进行导引，促使大脑放松，从而调节植物神经系统的机能，并在大脑放松后采用一定的套语振奋精神，进行自我动员的心理训练方法。

自我暗示和放松训练的放松部分可使大脑放松。消除心

理紧张，解除疲劳、加深睡眠，提高人体工作能力。自我暗示和放松训练可使人出现精神振奋和高度积极的状态，有利于参加考试和取得优异考试成绩。

自我暗示和放松训练可选择安静、昏暗的环境，采用坐势或仰卧姿势每天定时进行。考试期间则可在考试前后进行。练习时首先调节呼吸、暗示语可用“我的呼吸是安静的”等。然后可以逐级放松肌肉和大脑。可由上向下或由下向上进行。暗示语可用“我的右手放松了”等等。接着是对内脏器官发生影响的练习，暗示语部分可采用套语暗示，如“我休息得很好，我感到气足力盛，我准备行动，很愿意参加面临的考试……”等等。应试者做完后休息片刻，便可以精力充沛地参加实考。

第五节　考试卫生

考试具有许多优秀的功能，但若不注意或不懂得考试卫生知识，就会使应试者身心健康受到不同程度的影响，造成不良的后果。因此，应试者懂得一些考试卫生知识是非常必要的。通常所讲的考试卫生包括三个方面的内容：生理卫生、心理卫生和环境卫生。

一、生理卫生

考试是一种紧张的脑力劳动，它需要旺盛的精力。在高级神经系统紧张的条件下，碳水化合物、脂类、维生素A和B族（B_1，B_2，B_6，PP）以及维生素C的代谢过程加强，但热量消耗不增加或只是稍微增加。因此，应试者在考试期间

的膳食，应着重增加蔬菜、水果、动物性食品和豆类食品，减少纯糖和纯油脂性食物。应试者在备考期间避免过度疲劳而影响食欲，导致身体状况下降。应试者的营养应符合下列要求：

1. 营养必须保证质量。

2. 每天的食量不可过多，但必须吃饱，食物要精美，易于消化。

3. 营养必须满足对维生素和矿物质的需要。

4. 营养必须全面，食物要美味可口，增进食欲，植物性食物和动物性食物要兼顾。

5. 应根据符合卫生要求的营养计划行事。

应试者应对影响考试卫生的行为与方式进行斗争，切忌嗜好品和服用不利于身心健康的药物。所谓嗜好品是指实际上不能产生热量，只是对味觉和嗅觉神经、肠胃活动、心脏循环系统或中枢神经系统产生某些作用的食品，例如咖啡、茶、可可、香料、含酒精的饮料和香烟等。由于酒精和尼古丁对身体特别有害，因此应试者应回避这些东西。酒精首先影响大脑皮层，病状的程度取决于血液中的酒精含量。酒精有麻醉作用，只不过是麻醉的范围较小。根据酒精的不同饮量，能引起控制不住自己的理智、饶舌、反应变慢，考试注意力不集中、精神错乱等症状。此外，饮酒也违背考试卫生规则以及应试者的伦理和道德行为。

吸烟对健康和考试成绩的危害特别大。尼古丁通过口腔、咽喉和肠的粘膜而被吸收，最后在肝脏里分解。它能使血压上升，皮肤血管收缩，从而使皮肤的温度下降。开始时使心脏冠状血管扩张，但接着却使该血管收缩。这一点危害考试

时调节机制，因此对应试者的危害尤为明显。据此必须向应试者讲清楚吸烟的害处。同时还应注意，在考场上要禁止抽烟。

由于应试者求胜心切，想通过“特别的”措施取得更好的考试成绩，因而实考那一天离开正常的备考日的常规而另搞一套不合卫生的东西，结果反而使成绩下降。怎样备考，就应该怎么实考。这是因为机体对惯用的训练系统已经适应，而对“特殊的”措施还不适应，因而容易卡壳，导致成绩下降。为了防止这种现象，不要以为多睡一点觉，吃一点新花样，补充一点特别的东西（如在考前大量服用葡萄糖）就能提高成绩，备考时合理的、卫生的生活方式，有计划的备考以及实考之后的相应行为，这三者是取得良好的考试成绩的前提。

二、心理卫生

所谓考试心理卫生，是保障考试时期应试者心理健康正常状态的措施和各种活动的总和。考试心理卫生的任务是针对应试者在备考和实考中的细微适应障碍和心理上的不安等及时进行处理和预防，以保持和促进应试者的心理健康。

大脑是人体的高级神经活动中枢，其基本功能单位是神经细胞，它具有兴奋和抑制两种功能。大脑任何部位的兴奋能力均有一定限度，超过它的限度，大脑的工作能力就会下降。不仅应试者的身心会受到损害，考试成绩也未必能反映应试者掌握知识的真实水平。因此，考试必须适度。应试者要妥善处理人际关系；解决好理想与实现的矛盾；加强思想修养；开展心理卫生咨询。

三、环境卫生

应试者考试期间，应保证有足够的休息和睡眠时间。尽量使自己在安静、光线充足的环境中备考。要注意阅读卫生，复习时每隔四十五——五十分钟要进行一定时间的休息。这样既可预防近视眼以及其它疾病，又便于消除疲劳。同时要适当参加体育锻炼，以保证有旺盛的精力投入紧张的备考。

附录　特殊型考试方法探讨

许多单位或部门为了某种目的，或为了完成某一项任务，需要对一些特殊的事件或人物（如象学生实习成绩，教师教学质量，车间管理质量等）进行考评。但由于事物的模糊性，给度量带来不少困难。为此，本章从模糊性与随机性角度作进一步的探索。

第一节　模糊综合考评方法

一九六五年，美国控制专家、加利福尼亚州立大学教授查德（L.A.Zaden）发表了《模糊集合》一文，推广古典集论。对已成为现代数学的基础的康托的集合论，要修改集合的概念，当然是一件破天荒的事。然而，此后的二十年来，办杂志、出专著、形成一支不小的队伍，创立了一个新兴的数学分支 — 模糊数学。不论在国际上还是国内，模糊数学的研究正方兴未艾。其应用已涉及自动控制，信息处理，无线电控制、天气预报、医疗诊断、人工智能、语言学、生态学和管理科学等多项领域。而在考试中，FuZZy数学以自身独特的思想方法，闯开了考评模糊事物的大门，给众多的事物的量化与考评提供了优秀的工具，使考试学的科学研究加速发展，为它带了灿烂的前景。

一、模糊变换

设U、V均为有限集，即

$$U=\{u_1, u_2, \cdots, u_m\}$$
$$V=\{v_1, v_2, \cdots, v_n\}$$

V上的模糊子集可表为n维向量

$$\underset{\sim}{A}=\mu_{1\underset{\sim}{A}}/u_1+\mu_{2\underset{\sim}{A}}/u_2+\cdots+\mu_{n\underset{\sim}{A}}/u_n$$

或

$$\underset{\sim}{A}=(\mu_{1\underset{\sim}{A}}, \mu_{2\underset{\sim}{A}}, \cdots, \mu_{n\underset{\sim}{A}}) \xlongequal{\text{简记为}} (a_1, a_2, \cdots, a_n)$$

同理U上的模糊关系$\underset{\sim}{B}$可写成

$$\underset{\sim}{B}=(\mu_{1\underset{\sim}{B}}, \mu_{2\underset{\sim}{B}}, \cdots, \mu_{m\underset{\sim}{B}}) \xlongequal{\quad\quad} (b_1, b_2, \cdots, b_m)$$

设$\underset{\sim}{R}$是从U到V的一个模糊关系

$$\underset{\sim}{R}=\begin{pmatrix} r_{11} & r_{12}\cdots & r_{1m} \\ r_{21} & \cdots\cdots\cdots & r_{2m} \\ \vdots & & \\ r_{n1} & \cdots\cdots\cdots & r_{nm} \end{pmatrix}_{n\times m}$$

则根据模糊矩阵的复合运算，由$\underset{\sim}{R}$确定了一个变换：任给V上的一个模糊子集$\underset{\sim}{A}$，便可确定U上一个模糊子集$\underset{\sim}{B}$，即

$$\underset{\sim}{B}=\underset{\sim}{A}\circ\underset{\sim}{R}$$

这就是所谓模糊变换。利用模糊变换作为工具，可以解决特殊类型事件所需要的综合评判的问题。

二、模糊综合考评的数学模型

我们可以利用模糊变换进行综合评判。因为事物往往具有多种属性，因此评价事物也要兼顾各个方面，特别是在生产规划、管理调度、社会经济等复杂系统中，要作出任何一个决策，都须对多个相关因素作综合考虑，这就是综合评判的问题。

设评语集合（即代表等级、分类等的集合）为

$$U=\{u_1, u_2, \cdots, u_m\}$$

共m个等级。因素集合为$V=\{v_1, v_2, \cdots, v_n\}$共n个因素。设第i个因素的单因素评判为$R_i=(r_{i1}, r_{i2}, \cdots, r_{im})$，它可以看作是U上的一个模糊子集，其中$r_{ik}$表示第i个因素的评判对于第k个等级的隶属度。n个因素的总的评判矩阵为

$$\underset{\sim}{R}=\begin{pmatrix}\underset{\sim}{R}_1\\ \underset{\sim}{R}_2\\ \vdots\\ \underset{\sim}{R}_n\end{pmatrix}=\begin{pmatrix}r_{11} & r_{12}\cdots & r_{1m}\\ r_{21} & \cdots\cdots & r_{2m}\\ \cdots & & \\ r_{n1} & \cdots\cdots & r_{nm}\end{pmatrix}_{n\times m}$$

在进行综合评判时，当然要考虑各个因素对评定等级所起作用的大小，这种评判作用就形成了因素集合V上的一个模糊子集$\underset{\sim}{A}$

$$\underset{\sim}{A}=(a_1, a_2, \cdots, a_n)$$

a_i为v_i对$\underset{\sim}{A}$的隶属度，它就是单独考虑因素v_i对评判等级所起作用大小的度量，代表了根据单因素v_i评判等级的能力。其数值只有根据经验判断给出。

给定$\underset{\sim}{A}$，$\underset{\sim}{R}$后，即可进行综合评判，这种运算一般写成如下形式

$$\underset{\sim}{B}=\underset{\sim}{A}*\underset{\sim}{R}$$

如用框图表示，则为

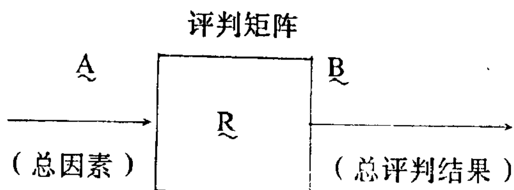


图14—1

根据对于$\underset{\sim}{A}*\underset{\sim}{R}$取不同类型的运算，就有不同的模型，今介绍几种常用的模型如下

模型Ⅰ　记作M（∧，∨）

取　$\underset{\sim}{A}*\underset{\sim}{R}=\underset{\sim}{A}\circ\underset{\sim}{R}$

即　$\underset{\sim}{B}=(b_1, b_2, \cdots, b_m)=\underset{\sim}{A}\circ\underset{\sim}{R}$

其中　$b_j=\bigvee\limits_{i=1}^{n}(a_i\wedge r_{ij})=\max\{\min(a_1, r_{1j}),$

$\min(a_2, r_{2j}), \cdots, \min(a_n, r_{nj})\}$

其物理意义是将原来的隶属度r_{ij}修正成$r_{ij}^*=a_i\wedge r_{ij}=\min(a_i, r_{ij})$，显然，在考虑多因素时$v_i$（$i=1, 2, \cdots, n$）的评判对任何等级$u_j$（$j=1, 2, \cdots, m$）的隶属度都不能超过$a_i$，同时对每个等级$u_j$而言（即决定$b_j$时）只考虑到那个起主要作用的因素，而未顾及其他因素的影响，这是一种“主因素决定型”的综合评判。

模型II　记作M（·，∨）

将模型I中的∧改成·，就变成模型II，即

$$b_j=\bigvee_{i=1}^{n}(a_ir_{ij})=\max\{a_1r_{1j},a_2r_{2j},\cdots,a_nr_{nj}\}$$

这里修正后的r_{ij}变成　$r_{ij}^*=a_ir_{ij}$，

事实上，模型II是一种“主因素突出型”的综合评判。

模型III　记作M（·，⊕）

⊕表示环和，其定义为

$$\alpha\oplus\beta=\min(1,\ \alpha+\beta),$$

显然环和不可能超过1。$\sum_{i=1}^{n}$表示对n个数在⊕运算下求和

$$b_j=\sum_{i=1}^{n}a_ir_{ij}=\min\{1,\ \sum_{i=1}^{n}a_ir_{ij}\}$$

模型III与模型II有两点重大的区别：

（1）在决定各因素的评判对等级u_j的隶属度b_j时，考虑了所有因素v_i（$i=1,2,\cdots,n$）的影响，而不是只考虑对b_j影响最大的那个因素；

（2）由于同时考虑所有因素，所以各a_i具有权系数的含义，因此a_i应满足归一化的条件，即$\sum_{i=1}^{r}a_i=1$。可以说，这个模型代表“加权平均型”的综合评判。

由于实际上$\sum_{i=1}^{n}a_ir_{ij}\leqslant1$，所以运算⊕蜕化为一般实数加法。因此M（·，⊕）＝M（·，＋），简单地说，模型III的物理意义是加权平均，数学运算变成一般矩阵乘法。

模型IV　记为M（∧，⊕），即

$$b_j=\min\{\sum_{i=1}^{n}\min(a_i,r_{ij})\}$$

这里并无$\sum_{i=1}^{n} a_i = 1$的要求。

以上是介绍了四种常用的综合评判的模型。对于同一评判对象，在同样的$\underset{\sim}{A}$，$\underset{\sim}{R}$下，按各种模型算得的结果$\underset{\sim}{B}$不同。这和人们从不同的角度观察同一事物而可能得出不同的结论一样。可以证明，其相对的大小顺序如下

$$\underset{\sim}{B}(\wedge, \oplus) \geqslant \underset{\sim}{B}(\cdot, \oplus) \geqslant \underset{\sim}{B}(\cdot, \vee)$$
$$\underset{\sim}{B}(\wedge, \oplus) \geqslant \underset{\sim}{B}(\wedge, \vee) \geqslant \underset{\sim}{B}(\cdot, \vee)$$

在实际应用中，综合评判的最后结果B的绝对大小没有多大意义，有意义的是不同对象间的比较，即相对大小。

给出一组事物，为了评判它们之间的优劣，可先用M（∧，∨）和M（·，⊕）计算，再在M（·，∨）和M（∧，⊕）中选择其一。由上述不等式，当算得的B（∧，∨）和B（·，⊕）的值偏小时，宜选M（∧，⊕）来计算，反之则选M（·，∨）计算。

为了综合不同的结果，可对不同的模型进行多层次的综合。

三、模糊多级综合考评的数学模型

在复杂系统中，需要考虑的因素往往很多，因素间还可能分属不同的层次。在遇到这类问题时，往往把因素集合按某些属性分成几类，先对每一类（因素较少）进行综合评判，再对评判结果进行各类之间的高层次的综合，例如我们要对某些对象进行评比，先分成几个大的方面，按上一节中的方法处理，显然这里每一方面的单因素评判又是低一层次的多因素综合评判的结果。同样，低一层次的单因素评判也可以

是更低层次的多因素评判的综合，为此提出了多层次（级）综合评判的模型。

给定集合V，若按下列原则将V分成n个子集：

$$\bigcup_{i=1}^{n} V_i = V$$

$$V_i \cap V_j = \phi \qquad i \neq j$$

在这种划分下得到的集合记为

$$V = \{ V_1, V_2, \cdots, V_n \}$$

多级综合评判可按下述步骤进行：

步骤 1　将因素集合V按上述原则划分成子集，作为第二级因素集合，即

$$V = \{ V_1, V_2, \cdots, V_n \}$$

记$V_i = \{ V_{i1}, V_{i2}, \cdots, V_{ik_i} \}$，$i = 1, 2, \cdots, n$。显然$V_i$含$k_i$个因素，V共有$\sum_{i=1}^{n} k_i$个因素

步骤 2　对每个V_i的k_i个因素，按初始模型作综合评判，设考虑V_i中诸因素在等级评判中所起作用的大小，得到V_i上的一个模糊子集$\underset{\sim}{A}_i$，V_i的总的评判矩阵为$\underset{\sim}{R}_i$，则得到

$$\underset{\sim}{A}_i * \underset{\sim}{R}_i = \underset{\sim}{B}_i = (b_{i1}, b_{i2}, \cdots, b_{im}) \quad i = 1, 2, \cdots, n$$

步骤 3　对V的n个因素按初始模型作综合评判，V_i的综合评判结果为$\underset{\sim}{B}_i$，它是V中单因素V_i的评判。设考虑V中诸因素所起作用的大小，得模糊子集$\underset{\sim}{A}$，总评判矩阵为

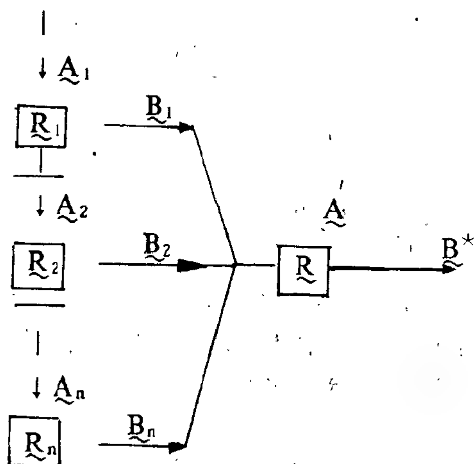


图2

$$\underset{\sim}{R} = \begin{pmatrix} \underset{\sim}{B}_1 \\ \underset{\sim}{B}_2 \\ \vdots \\ \underset{\sim}{B}_n \end{pmatrix} = (b_{ij})_{n\times m}$$

于是得到$\underset{\sim}{B}^* = \underset{\sim}{A} * \underset{\sim}{R}$，它既是V的综合评判结果，也是V的所有因素的综合评判结果，可写成算式如下

$$\underset{\sim}{B}^* = \underset{\sim}{A} * \underset{\sim}{R} = \underset{\sim}{A} * \begin{pmatrix} \underset{\sim}{A}_1 * \underset{\sim}{R}_1 \\ \underset{\sim}{A}_2 * \underset{\sim}{R}_2 \\ \vdots \\ \underset{\sim}{A}_n * \underset{\sim}{R}_n \end{pmatrix}$$

框图如图2所示。

这就是所谓二级模型。当然，必要时还可继续划分，得到三级乃至更多级综合评判模型。

二级模型既反映了客观事物因素间的不同层次，又可避免由于因素过多而难于一下子确定各因素在评判中所起作用大小的模糊子集A的隶属函数的问题。

第二节　修正模糊综合考评

修正模糊综合考评是力图在模糊综合考评的基础上，定义并应用“一致性”的概念，把反应某一事物的信息进行综合处理，提出修正模糊综合考评方法的数学模型。下面就教学质量的考评这一具体事件来，阐述这一思想与方法。

对教学质量进行评估，是重视教学，调动教师积极性的一项重要措施。同时，通过评估，可比较全面的了解教学的现状，教师的教学水平，存在哪些问题，这就是鼓励先进，鞭策后进，造成某一群体以外的压力，使教师不断地改革教学方法，更新教学内容，促进教学质量的提高，但是教学质量评估是一个极其复杂的问题，搞得好能调动教师的积极性，促进教学质量的提高，搞得不好会产生副作用，影响教师积极性的发挥。为了搞好教学质量的评估，除了制定一些原则，如象目的性原则，可比性原则，客观性原则，接受性原则等外，对评估的技术方法的研究是非常有意义的。为此，本节提出教学质量的修正模糊综合评价法。

一、教学质量的分级系统与指标

教学质量是一个总的概念，它的优劣不是一个单因素的问题，对其评价往往涉及多种因素，当前教学质量评估，已经引起大家的极大重视与兴趣，发表了不少有见解的论文。

根据“教育面向现代化、面向世界、面向未来”的思想，结合实际工作中的体会，我们认为衡量教师教学质量高低的评估标准主要有四方面的内容：

（一）（F_1）教学能力方面，主要包括：

1.（F_{11}）按照教学大纲的要求，教师在选用教材与参考书方面，是否具有思想性，先进性，科学性，逻辑性；教材是自编，改编，还是采用统编教材，自编、改编的教材水平如何？能否注意课程的衔接。

2.（F_{12}）教师授课能否完整地掌握教材内容，突出重点、难点，能否吸收最新科研成果，并能否抓住教材的主要内容选择、布置习题，拟订考试题。

3.（F_{13}）教师在教学方法上，能否理论联系实际；能否运用最新的教学手段，用启发式进行教学；能否因材施教，循循善诱；能否引导学生积极思维，独立思考，培养学生运用知识的能力。

4.（F_{14}）教师能否既教书又育人，言传身教，关心学生思想政治上的成长。

（二）（F_2）业务水平方面，主要包括：

1.（F_{21}）教师对本门学科是否具有扎实的基础理论、基本知识和实际操作的本领，是否了解并运用本学科国内外最新科研成果。

2.（F_{22}）教师担任研究生课程及本科一门或多门以上课程教学的能力；指导学生撰写论文及毕业设计的能力。

3.（F_{23}）教师在本门学科领域内进行科研的能力。

4.（F_{24}）教师的外语水平。

（三）（F_3）教学态度方面，主要包括：

1.（F_{31}）教师对待教学工作有无高度的事业心和责任感，能否做到严谨治学，认真负责地对待教学工作。

2.（F_{32}）教书育人，严格教学纪律，积极进行精神文明的建设。

3.（F_{33}）对学生因才施教，循循善诱，不抱偏见，严格要求，不感情用事。

4.（F_{34}）能否虚心征求和听取同行与学生对自己教学工作的意见，并不断改革自己的教学工作；与同行共同搞教学、科研工作时的团结协作精神与表现。

（四）（F_4）教学效果方面，主要包括：

1.（F_{41}）学生对大纲规定的基本理论，基本知识技能的掌握程度和灵活运用能力。

2.（F_{42}）学生是否具有独立思考，逻辑思维、分析、解答问题的能力和自学的能力；学生口头和书面表达能力的情况。

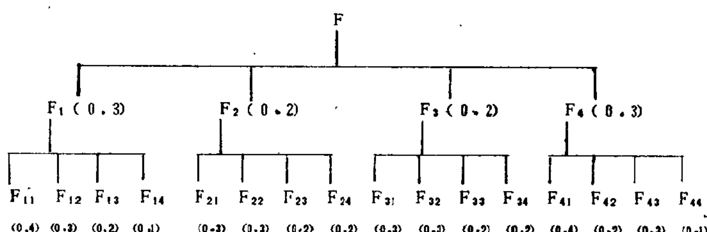
3.（F_{43}）学生是否具备勤奋好学、虚心求教，乐于助人的良好学习风气及品德。

4.（F_{44}）学生是否遵守学习纪律及教学秩序，考试不作弊。

以上四个方面既互相独立，又互相影响；实际应用有所侧重。但评估标准一经确定下来，我们就必须按这个统一的标准去做，不能随意降低标准，当然各自在标准掌握上会有所差异，这可以在后面的质量分析中加以解决。标准定下来以后，应稳定一个时期，以有利于保持评估连续性和可比性。

对一个复杂的问题，每个因素在综合评价中的地位和重要性是各不相同的。为了得到正确的评价结果，必须确定每

一层次诸因素的相对优先权重，即对每一个因素给出一个权数，确定权数可用几种方法进行。如层次分析法，是先将复杂的对象分解为一个分级系统，然后由领导和专家慎重地对同层次的诸因素进行两两比较，找出它们的优先顺序；又如特尔斐；是先设计好权重调查表，让不同类型的专家和有经验的教师在彼此隔离的情况下填写各因素的重要性程度，然后汇总整理，进行统计归纳，并对原先设计的权数进行修正，再反馈给专家。如此多次反馈（一般经过3～4轮），就可取得比较集中的意见，从而确定出各因素的权重。本文的权重如下：



二、教学质量考评方法

长期以来，我们所采用的考评教学质量的方法，主要是组织教学检查，它包括：学期测验考试，检查学生作业和教师备课；教师互相观摩听课，召开一些教师学生座谈会，听取教师、学生的意见等等。这些方法灵活且有弹性，无疑对了解教师的教学情况，交流教师的教学经验，不断提高教学质量均起到一定的良好作用，这也是我们今后仍将适当采用的方法。同时我们也应该看到这些方法仅仅囿于以定性上对教师教学质量进行考评，考评的准确性和效果受个人主观因

素的影响较大，而且还缺乏量的标准和比较，以至考评结果往往带有一定的局限性，主观性和片面性。要改变这种“传统型”的考评方法，就必须在我们思想认识上来一个根本的突破。伟大革命导师马克思曾指出：“一种科学只有在成功地应用数学时才算达到了真正完美的地步”。近年来，现代数学的分支——模糊数学在教育学科中的广泛应用，为客观考评教学质量提供了比较可靠的数量依据，它既注意了教学工作的灵活性，又克服了评估中的随意性，有着广阔的发展前景。

（一）模糊信息的获得

模糊综合评判法是建立在调查统计的基础上的，从广泛的调查中获得大量的信息。这些信息之所以是模糊的，是因为很多事物很难用确切的量来衡量。众所周知，我们很难象对学生的学习成绩那样，给教师的教学质量“打分数”，也说不上75分与76分之间有什么质的差别。因此在调查时，只宜比较笼统地评定“优”、“良”、“中”、“一般”、“差”。当然不同的人对其评价不尽相同。重要的是根据获得的大量模糊信息，用恰当的数学方法进行推理和运算。其结论不应是简单的优与劣，而应是在多大程度上为优，又在多大程度上为劣，这就是模糊推理。

为了获得这些模糊信息，应特别设计一份“教学质量调查表”，分别请有关领导、同行教师和学生用不记名的方式进行评价。每个因素按五个等级评价，即优、良、中、一般、差，请评价人在相应的格子内划“∨”记号。

（二）模糊综合评价的方法和步骤

在复杂系统中，需要考虑的因素往往很多，因素间还可

能分别属不同层次。先对一类（因素较少）进行综合评判，再对评判结果进行各类之间的高层次的综合。我们首先计算R。

$$R=\begin{bmatrix} R_1 \\ R_2 \\ R_3 \\ R_4 \end{bmatrix}=\begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \\ A_{41} & A_{42} & A_{43} \end{bmatrix}\begin{bmatrix} \frac{20\%}{n_1} \\ \frac{45\%}{n_2} \\ \frac{35\%}{n_3} \end{bmatrix}$$

其中A_{ij}表示，学生同行和领导对被评教师的教学主因素的评定矩阵。i与F_i相对应，j = 1、2、3与学生、同行和领导相对应，A_{ij}的列表示等级，行表示Fl的子因素的评定信息。n_1、n_2、n_3分别表示学生、同行和领导的人数。

其次计算分类模糊评判结果：

$$\underset{\sim}{F}_i=\underset{\sim}{A}_i \circ \underset{\sim}{R}_i=(f_{i1},\ f_{i2},\ f_{i3},\ f_{i4},\ f_{i5})$$

f_{ij}表示Fl主因素评定结果，j = 1、2、……5，对应于评定等级优、良、中、一般、差。

$$\underset{\sim}{A}_i=(F_{i1}\quad F_{i2}\quad F_{i3}\quad F_{i4})$$

$$\underset{\sim}{R}_i=(r^{[i]}_{kj})_{4\times 5}\quad k=1,2,3,4\quad j=1,2,3,4\quad i=1,2,3,4$$

运算○是“主因素决定型”，其运算按（∨，∧）进行。

$$f_{ji}=\max\{\ \min(F_{i1},\ r^{[i]}_{1j}),\ \min(F_{i2},\ r^{[i]}_{2j})\cdots\cdots$$

$$\min(F_{i4},\ r^{[i]}_{4j})\}$$

归一化：

$$\underset{\sim}{F}_i = \left[\frac{f_{i1}}{\sum_{i=1}^{5} f_{ij}} \quad \frac{f_{i2}}{\sum_{i=1}^{5} f_{ij}} \quad \frac{f_{i3}}{\sum_{i=1}^{5} f_{ij}} \quad \frac{f_{i4}}{\sum_{i=1}^{5} f_{ij}} \quad \frac{f_{i5}}{\sum_{i=1}^{5} f_{ij}} \right]$$

再次计算高一层次的模糊综合评判结果：

$$\underset{\sim}{F} = \underset{\sim}{B} \circ \begin{pmatrix} \overline{\underset{\sim}{F}}_1 \\ \overline{\underset{\sim}{F}}_2 \\ \overline{\underset{\sim}{F}}_3 \\ \overline{\underset{\sim}{F}}_4 \end{pmatrix}$$

$\underset{\sim}{B} = (F_1 \quad F_2 \quad F_3 \quad F_4)$　　（运算○同上）

fi表示教学质量的评判结果，i＝1，2，……，5对应于评定等级优、良、中、一般、差。

归一化：

$$\overline{F}_1 = \left[\frac{f_1}{\sum_{i=1}^{5} f_i} \quad \frac{f_2}{\sum_{i=1}^{5} f_i} \quad \cdots\cdots \quad \frac{f_5}{\sum_{i=1}^{5} f_i} \right]$$

其等级对应关系：$\frac{f_1}{\sum_{i=1}^{5} f_i} \quad \frac{f_2}{\sum_{i=1}^{5} f_i} \quad \frac{f_3}{\sum_{i=1}^{5} f_i} \quad \frac{f_4}{\sum_{i=1}^{5} f_i} \quad \frac{f_5}{\sum_{i=1}^{5} f_i}$

优　　良　　中　一般　差

根据最大隶属度判别准则：

$$\frac{f_1}{\sum_{i=1}^{5} f_i}, \quad \frac{f_2}{\sum_{i=1}^{5} f_i}, \quad \frac{f_3}{\sum_{i=1}^{5} f_i}, \quad \frac{f_4}{\sum_{i=1}^{5} f_i}, \quad \frac{f_5}{\sum_{i=1}^{5} f_i}$$

中最大值所对应的哪一等级，就为评定等级。但这种方法会丢掉很多有用的信息，特别是意见分散的情况。如F＝（0.2，0.19，0.21，0.2，0.2）按此方法可以评定为中，也可评为优，一般、差。因此，下面将讨论对评判结果的修正方法。

（三）修正方法

为了确定评定的意见分散与否，我们借用热力学的名词“相对熵”的概念，来定义“不一致”性的度量。

定义：设评判结果矩阵F＝（f_1，f_2，……f_n），则不一致H。

$$H=\frac{\sum_{i=1}^{n} f_i \ln f_i}{-\ln n}$$

（注意：$o\cdot \ln o=\lim\limits_{x\to 0} x\ln x=\lim\limits_{x\to 0}\frac{\ln x}{\frac{1}{x}}=\lim(-x)\doteq 0$）

可以证明H是0—→1之间的一个数，若H＝0，表示“不一致”程度为0，即完全一致；若H＝1，则表示“不一致”程度为1＝100％，即“完全不一致”，举一个“完全一致”的例子。

设F＝（0 1 0 0 0）此时100％的人意见都是良。

$$H=\frac{0\ln 0+1\ln 1+0\ln 0+0\ln 0+0\ln 0}{-\ln 5}=0$$

即“不一致程度”为0，也就是说“完全一致”，结论意见当然是良。

再举一个“完全不一致”的例子。

$$\underset{\sim}{F}=(0.2\quad 0.2\quad 0.2\quad 0.2\quad 0.2)$$

即意见平均分配从优到差全部五个等级，意见“完全不

致。

此时：$H = \frac{5 \times 0.2 \ln 0.2}{\ln 0.2} = 1$

即"不一致"程度为100%，我们根据试验确定H大于等于0.90即可结论为"意见分散"，通过许多的试验得出，根据H的值及峰值曲线的分布情况按以下四条原则给出结论等级。

（1）如果最高峰的0.618倍大于其他各峰值且$H<0.9$，则以最高峰的等级作为结论意见，最高峰值可作为模糊评判的从属，附在等级之后。

例　F=（0.7，0.2，0.05，0.05，0）

$0.618 \times 0.7 = 0.4326 > 0.2$及$0.05$

$H = 0.541 < 0.90$

结论：优（0.7）。

（2）若最高峰的0.618倍不大于次高峰值，但$H<0.90$，结论意见可定为最高峰等级另加符号"＋"及"－"。若次高峰在相邻的左边加"＋"号，在右边加"－"号，从属度按下式计算。

$$从属度 = 最高峰值 + \frac{次高峰值}{2}$$

关于从属度的计算可以这样解释：我们动员同意评为次高峰等级的人，一半改变意见为同意最高峰等级，为表示这种改变另加符号"＋"或"－"。

例　$\underset{\sim}{F}$=（0.11，0.53，0.33，0.03，0）

$0.618 \times 0.53 = 0.32754 < 0.33$

$H=0.652<0.90$

从属度$=0.53+\frac{0.33}{2}=0.7$，结论：良$^{-}$（0.7）。

（3）若最高峰的0.618倍不大于次高峰而次高峰与最高峰不相邻，此时即使$H<0.90$，也应结论意见分散。

例　$\underset{\sim}{F}=(0.4,\ 0.26,\ 0.34,\ 0,\ 0)$

$0.618\times0.4=0.2472<0.34$

$H=0.673<0.90$

但0.34与0.4不相邻，结论意见分散。

（4）$H>0.90$。结论：意见分散。

例　$\underset{\sim}{F}=(0.21,\ 0.19,\ 0.20,\ 0.20,\ 0.20)$

$$H=\frac{0.21\ln0.21+0.19\ln0.19+0.20\ln0.2+0.2\ln0.2+0.2\ln0.2}{\ln0.2}$$

$=0.999>0.90$

对各方面信息不一致，应进行深入的调查研究，找出各结论不一致的具体原因，从实际出发进行分析，调查、重评，最后判定总评等级。

教学质量修正模糊综合考评法，是我们在考评的数学模型的研究中，结合实际工作产生出来的。它有一定的实践基础与实用的可能性。当然本文所提出的权重数是否能反应所有的情况应用效果如何均有待于进一步实践与探索，也有待于考评的数学模型的研究发展来验证。但是就目前来看，这种方法是较科学的，能全面应用和处理反馈信息的优化数学模型。其不足之处是计算工作烦杂。但只要使用计算机进行反馈信息（数据）的处理，其效率是会受欢迎的。

第三节　齐次马尔可夫链分析法在教学效果考评中的应用

根据教学活动的特点和教学考评的需要，本文应用马尔可夫链的原理于教学考评过程，形成了齐次马尔可夫链评定法。此方法着眼于教学过程，重视"历史"，与其它教学效果考评方法来讲更符合具有紧密的前后联系的教学过程的实际情况，具有独自的特点。

教学效果的评定是教学质量考评的一个重要的组成部分。它是教学质量考评的核心内容。多年来对教学效果评定的研究，仍处在终结性考评方面，对教学效果形成性考评研究不足。为此，本节企图应用齐次马尔可夫链评定方法，对教学效果进行评定。

一、齐次马尔可夫链的概念

考虑一个可能具有m个状态（即1，2，…，m）的系统，状态的变化只发生在参数的离散值上。例如，在时刻t_1，t_2，…，t_n。令X_{n+1}表示在t_{n+1}时的系统状态。一般说来，系统将来处于状态i的概率与它的全部历史有关，所以应该用条件概率：

$$P(X_{n+1}=i|\ X_0=x_0,\ X_1=x_1,\ \cdots,\ X_n=x_n) \qquad (1)$$

表示，其中$X_0=x_0$，…，$X_n=x_n$代表系统以前所有的状态。如果将来的状态只与现在的状态有关，条件概率（1）变为：

$$P(X_{n+1}=j \mid X_0=x_0, \cdots, X_n=x_n)$$
$$=P(X_{n+1}=j \mid X_n=x_n) \qquad (2)$$

这样的过程称为马尔可夫链。

对马尔可夫链，可以将从时刻t_m的状态i变为时刻t_n的状态j的条件概率表示为：

$$p_{ij}(m,n) \triangleq P(X_n=j \mid X_m=i) \quad n>m \qquad (3)$$

如果$p_{ij}(m,n)$只有时间差t_n-t_m有关，而与时间起点t_m无关，则称该马尔可夫链为齐次的。此时，定义：

$$p_{ij}(k) \triangleq P(X_k=j \mid X_0=i)$$
$$=P(X_{s+k}=j \mid X_s=i) \quad S\geqslant 0$$

为k步转移概率。它表示齐次马尔可夫链以状态i经过k次转移之后到达状态j的条件概率。

对齐次马尔可夫链，设系统的初始状态可以由行矩阵表示：

$$P(0) \triangleq [p_1(0)\,p_2(0)\cdots p_m(0)]$$

式中$p_i(0)$是系统最初处于状态i的概率。经一步转移之后，系统处于状态j的概率由全概率公式给出：

$$p_j(1) \triangleq P(X_1=j) = \sum_i P(X_0=i)P(X_1=j \mid X_0=i)$$

亦即 $p_j(1) = \sum_i p_i(0)\,p_{ij}$

故可用矩阵表示为 $P(1)=P(0)P$

$$P \triangleq (p_{ij})_{m\times m}$$

P（1）也是一个行矩阵。

类似地，经过两步转移之后，系统处于状态j的概率为

$$p_j(2) \triangleq \sum_k P(X_1=k)P(X_2=j|\ X_1=k)$$

$$=\sum_k p_k(1)p_{kj}$$

或表为 $P(2)=P(1)P=P(0)PP=P(0)P^2$

同理可以证明，经过n步转移之后，系统状态概率矩阵为

$$P(n)=P(n-1)P=P(n-2)PP=\cdots=P(0)P^n$$

由上可见，系统在任何时刻的状态概率是由初始状态概率和转移概率确定的。

在数学上，齐次马氏链有一重要的性质：当$n\to\infty$时，P（n）的极限是齐次马氏链在平稳状态下的概率分布，即系统在客观上不再发生变化。

$$\lim_{n\to\infty}P(n)=\lim_{n\to\infty}P(n+1)=P(\text{平衡})$$

二、齐次马尔可夫链评定法

在教学效果评定的量化中，齐次马尔可夫链评定法把一个集体（一个班级或一个年级）中学生获得优、良、中、及格以及不及格各种等级学生人数占总人数的比例作为状态变量，并用向量P（t）表示之，

$$P(t)=[X_1(t)\ \ X_2(t)\ \ X_3(t)\ \ X_4(t)\ \ X_5(t)]$$

比如，经一次考试，一个班级60名学生中，获得优等的

学生有11人，良等的有20人，中等的有17人，及格的有8人，不及格的4人，那么，状态向量就可写成：

$$P(t)=[11/60 \quad 20/60 \quad 17/60 \quad 8/60 \quad 4/60]$$
$$=[0.18 \quad 0.33 \quad 0.28 \quad 0.14 \quad 0.07]$$

t表示时刻，在此只取1，2，……等整数值。

这一向量一般称为初始向量，教学效果的齐次马氏链评定法还需要了解通过一周期教学后的下次考试中上述各等级学生的变化情况。拿上面的例子，在第二次考试，原来获得优等成绩的11名学生继续保持优等的有7人，下降为良的有2人，下降为中的2人，其他为0人。由此我们得到了第一次考试中获得优等成绩的11名学生的转移情况：

$$[7/11 \quad 2/11 \quad 2/11 \quad 0 \quad 0]$$

其余良、中、及格、不及格等级的情况如下：

$$[3/20 \quad 10/20 \quad 7/20 \quad 0 \quad 0]$$
$$[1/17 \quad 4/17 \quad 7/17 \quad 4/17 \quad 1/17]$$
$$[0 \quad 1/8 \quad 2/8 \quad 4/8 \quad 1/8]$$
$$[0 \quad 0 \quad 1/4 \quad 1/4 \quad 2/4]$$

这一变化用矩阵来表示就有：

$$P=\begin{bmatrix} 7/11 & 2/11 & 2/11 & 0 & 0 \\ 3/20 & 10/20 & 7/20 & 0 & 0 \\ 1/17 & 4/17 & 7/17 & 4/17 & 1/17 \\ 0 & 1/8 & 2/8 & 4/8 & 1/8 \\ 0 & 0 & 1/4 & 1/4 & 2/4 \end{bmatrix}$$

P称为转移矩阵，不难看出，在P(2)与P(1)之间成立：

$$P(2)=P(1)\cdot P$$

$$=[11/60 \quad 20/60 \quad 17/60 \quad 8/60 \quad 4/60]\begin{bmatrix}7/11 & 2/11 & 2/11 & 0 & 0\\ 3/20 & 10/20 & 7/20 & 0 & 0\\ 1/17 & 4/17 & 7/17 & 4/17 & 1/17\\ 0 & 1/8 & 2/8 & 4/8 & 1/8\\ 0 & 0 & 1/4 & 1/4 & 2/4\end{bmatrix}$$

$$=[0.18 \quad 0.28 \quad 0.32 \quad 0.15 \quad 0.07]$$

关于教学系统状态转移矩阵P极限向量的求法，为了实用的方便，我们给出下列步骤。

1. 列出学生个体成绩等级转移表。转移表除学生序号外的共有三列，第一列为该生初始状态的考试的成绩；第二列为未始状态的考试成绩；第三列表明该生在初未两次考试中成绩转移的情况，ij表明该生从i等转向j等。

2. 确定转移矩阵P，转移矩阵P中元素P_{ij}由下式决定：

$$p_{ij}=\frac{ij\text{的频数}}{\sum_{j=1}^{5}[ij\text{的频率}]}$$

$$P=(p_{ij})_{5\times 5}$$

3. 求出转移矩阵的逆矩P^T的极限向量。在数学上，求转移阵的极限向量即为求矩阵P^T的特征值为1的特征向量，它包括下列几个方面的工作：

（1）求出矩阵$G=I-P^T$

（2）列出特征方程$G\cdot X=0$

式中　$X^T = (X_1\ X_2\ X_3\ X_4\ X_5)$，加上约束条件$\sum_{i=1}^{5} x_i = 1$，得方程组

$$G \cdot X = 0$$

$$\sum_{i=1}^{5} X_i = 1$$

（4）我们记X为特征值为1特征向量，解出向量$X = [x_1 x_2 \cdots x_n]$，这一向量即为G的极限向量，根据最大项原则，可用其中值最大的一等级表示教学工作的效果。

第四节　考试的展望

一、客观测量的兴起

本世纪初，世界各国采用的是主观性考试。但是，20年代直至第一次世界大战前后，在西方（特别美国），为了科学地选拔人才，甄别人才，与主观性考试所不相同的种种标准化智力测验，乃风起云涌，测验专家大量编制一些客观性试题，经过大量试测，其有效性和客观性业已证明达到一定数量指标，然后发行供给人们使用，这时代的变化，大大促进了教育测量这门科学的建立和发展。当时，这种测验就是后来客观性考试的原型。它的特点是试题的取样亦即智能的复盖面广、效度高，题式答法单纯，不要求考生作出长篇大论，评分客观，准确；命题、考试实施到评定成绩，努力排除一切无关因素的影响，并实施考试条件的规范化，以确保得分的准确性，可靠性和可比性；每一试卷保持同效的衡量

指标；每考试都根据对全国范围或大规模的同年龄、同年级或同性质的学生集体进行实测，据以制成集体常模，借助于这些常模，任何考试在测验中所得的原始分数，就可以转换成常模性的分数。

客观性考试的蓬勃发展，为教育工作者提供了其客观性和正确性在某些方面甚与物质量具相比拟的教育量具，大大促进了教育研究的科学水平，也大大丰富了教育测量学的理论，但是经过客观性处理的各科教育测验，其品种毕竟是有限的。为了满足各种不同阶段，不同单元教学上的需要，教师自编的课堂测验仍然贯居于不可缺的首要地位，经过彻底改革的课堂考试，除了无须象客观测验那样要经过试测，计算出常模，制备覆份以及采取一系列的标准化的手续之外，都是要求教师按照标准测验编制的主导思想与原则方法进行编制的，其性质、试题取样、试题形式、实施方法以至评分所根据的原则与标准测验都大体相同。应当说，20年代到40年代这样的悠久的历史阶段中所有一切心理与教育测验都是以鉴别考生的个别差异为指导思想的，成绩评定亦是以个别差异的实际分布作为依据的，因而这一历史时期的各种测验往往称为“常模参考性”考试，其评分方法则属于“相对评分法”的范畴。

二、目标参考性考试与绝对评分法的提倡

常模参考性考试的发展，使成绩考评实现了高度的数量化，定分的客观性与可比性达到了前所未有的水平，这是其最大的贡献所在，但是由于这种测验的立足点是个别差异，是比较不同考生在学业上的“总”的成就，具有一般性的调

查性质，而不是着眼于首先明确规定一个测验所要考查的各项目标、不是着眼于鉴别各类教学目标是否完成或完成得一样好。它对考生学习所起的作用主要是考核或监督的功能，而不能充分起到诊断学习优劣、难易的，主动调节努力方向，确保完成各项学习目标的作用。60年代后，在智力群因素结构新理论和其它新教学论的影响以及课程改革迫切需要的推动下，一种与常模参考性考试相对而立，被称之为“目标参考性考试”的新型测验乃应运而生。

目标参考性考试与常模参考性考试相比较，前者有较多的优势。第一，目标参考性考试的用处在于确知有哪一些规定的教学目标某一学生经已完成，而后者的用处则在于确知某一学生对于某科目的知识量掌握了多少。由实施目标参考性考试而制作的初步成绩记录往往是标明一系列业已完成或尚未完成的学习目标，而常模参考性考试后的初步成绩记录，则是全部考试试题中已被答对的总计题数。第二，目标参考性考试的“目标”就是完成所有的教学目标。考生学习的好坏是以该生对预定的各项教学目标已完成的数量或百分数来判定的，而常模参考性考试的“常模”是指某一规定的学生集体在该考试的成就。某一特定考生的学习成就的好坏，乃是以该生成绩在这一规定集体所居地位如何作为判断的。

在欧美现代教学体制的许多革新中已对目标参考性考试广泛地应用，尤其是需要严格贯彻循序渐进的、依靠自学或自动调节进度的各种教学体制包括计算机辅助，计算机管理的教学体制中，其应用的效果更为显著。在所有这些教学体制中，测验总是与教学整合为一的，在单元教学之前，中期和结束，都必须通过考试来核对所必须具备的基础知识技能，

诊断可能出现的学习困难，并预订后继的教学程序。

三、电脑在客观性考试中的应用

电脑的应用，是使考试标准化、科学化的一个重要技术手段。它主要表现在以下几方面：

第一，试题的分析。在试题编制的过程中，电脑主要是提供试题的各项质量指标（难度、区分度等），供编题人员分析答考。即于试题初步编毕试测后，用电脑对每一道题进行项目分析，根据电脑分析得出的数据，对题目进行修改和调查。分析结果用电脑打印成绩表格的形式。

从表中可清楚地看出每一道路的答对率，难度指标，区分度指标，选各选择支的人数及其平均分。据此表，命题人员即可把区分度差的或难度不合适的题目修改或撤换，而达到要求的题目就可以按照难易，类型分别存在题库中备用。

第二，试卷的编制。题库由电脑建立，保密性强，抽取也方便，当要编制试卷时，由电脑按照类型与难易随机抽取即可。一份试卷中题目的难易程度应如何分布才适合呢？实践证明：以下的题目难易程度分布比例可使试卷具有较高的信度和区分度。

难易程度	最难	较难	适中	较易	最易
分　　布	5%	15%	60%	15%	5%

一份试卷经试测并由电脑进行项目分析后，就可以得到这份试卷的题目难易程度以及区分度的分布情况，以作为从整个试卷编制角度来考虑取舍题目的依据。

第三，考试的组织。考试的组织很多属于行政管理上的工作，如用电脑安排考场，分配考号，核对准考证，登记分数，印发成绩通知书等。由于考生以及考卷的所有信息都存在电脑里，进行这些工作对电脑来说是轻而易举的。

第四，评卷与统计。选择题的试卷由电脑评改统计，是十分简单的，即将正确答案和考生的考卷分别输入电脑，即可根据要求得到每个考生的各大题分数及总分，用光电阅读器向电脑中输入考卷，每小时可评阅几千以至几万份卷子。既节约了大量人力和资金，又大大地减少了差错。分数表打印出来后，电脑可以随即按总分高低排序打印出另一个分数表及一系列统计数据。如平均分，均方差，信度，测量误差，分数分布图以及分布峰值，偏整指标等。统计出来的数据可与别的考试相比较。通过求整个试卷各个题目的相关系数矩阵以及因素分析矩阵，还可以了解到一份试卷中，包括有多少个主要因素，即考了些什么，主要体现什么能力。并且可以了解这些因素在试题中的分布情况，是否与命题时所设想的一致，以验证命题时的意图与原则一致。

第五，分数的转换。一个标准化考试，试题即使经过了预测，修改和调整，到真正考试后，其实际结果与原设想还是有一定的差别的。这就要求进行分数转换。所谓分数转换就是将原始分数按一定常模转换成标准分。使它具有等值性和可比性。等值就是说相同水平的考生无论何时何地考，其所得的分数应相同。而可比就是说同一标准化考试，不同的试卷，即使难易上有差别，但所得的分值是可以进行比较的。转换的方法有许多种，如此两个试卷中相同的题目（称为平衡题）的分数为标准进行转换，这种转换比较可靠，但要求

考过的试卷要保密。又如设立观察点，这些观察点的考生必须能稳定地反映整体考生的水平，统计后所得的原始分以观察点的分数为准，按常模转换。两种方法的转换都由电脑直接自动进行，先由电脑算出平衡题的分数或观察点的分数，然后对整体的成绩进行标准化，它包括将主观性试题分数进行的调整。向考生公布的成绩，就是已经进行标准化转换后的分数。总之，在客观化考试中，电脑的使用是必不可少的，前景是无量的。

四、考试的作用和局限

世界上许多国家都在惊叹学校教育陷人“考试主义的泥潭”，因此，锐意改革是一个潮流。大致有如下三个趋势：第一，由单纯用考试测验鉴定教育效果向教育质量的全面评价转化　。因为考生的态度、行为、创造力和学习方法等很难用考试加以鉴定，必利用各种手段综合考察教学效果。第二，由竟争、选拔模式的考试向促使学生个性发展的考试模式转化．这种考试只把考试成绩与确定了的目标相比较，不排名次，不列等级，不与他作横向比较，以减轻学生的心理负担，促进个性的全面发展。第三，东、西的考试制度的渗透与合流。西方教育界强调客观的，标准的考试，而东方则重视对教育目标进行综合评价的五级记分制，近来两者开始相互借鉴与渗透。

学校教育陷人“考试主义泥潭”，整个社会中也有在逐步陷入考试的深渊的，考试的分数每天都在使人上“天堂”，同时也在使人下“地狱”，考试为何有如此的力量呢？是人们对它有宗教式的信仰，还是它本身就完美无缺，是出自于

愚昧，还是出自于科学。我们不得不深思。想必要对有些问题进行思考。

分数是评价考生的唯一尺度吗？这个问题已讲得很清楚。而现实中的分数就有其特殊的影响。不少家长和社会人士，常把考试分数的偏低视为考生懒惰、懈怠和不够用心的标志，把分数绝对化，其实这是大大的费解，苏联当代杰出的教育家瓦·阿·苏霍姆林斯基曾指出："学习、上课、完成作业，经常得到分数——这一切绝对不应当成为用来衡量、评价一个人的唯一的、概括一切的尺度。学生是年龄尚小的人，他对这种日常的衡量的体会和感受，特别敏锐和极其脆弱。应当使学生通过亲身体验，深信人们是用许多尺度衡量他们的，是从各个方面来看待他的。"青少年如此，成人同样也是如此，对成年人的考试的分数常与能力相差甚大。当然，这绝不是说考试无用，只是证明分数不是全面评价考生的唯一尺度。目前高等学校招生就是考试与推荐相结合。

另一方面，分数不是反映考生在学习上取得成就的唯一标准。苏霍姆林斯基认为："学习上的成就这个概念本身就是一种相对的东西。对一个学生来说，'五分'是成就的标志，而另一个学生来说，'三分'就是了不起的成就"。"教师在给学生做出评价的时候，不要只评价所谓纯粹的能力，而要评价劳动和能力的统一物，并且把劳动放在首位，只有这样的评价在道德方面才是正确的"。很简单，学习上取得的成就不是绝对分数所能刻划的，它必须通过处理，或找其学习前后分数差值的合理刻划因素进行描述。

考试是检验人的知识和智能的一种较有效的工具。是检验人的知识和智能的最终标准——实践的一种中介。考试与

社会以及社会发展之关系是有待于人们去讨论和研究的 它的科学化、现代化是考试发展的必然。